

# Studies In Aspects of Peer Review: Novelty, Scope, Research Lineage, Review Significance, and Peer Review Outcome

Submitted in partial fulfillment of the requirements  
for the degree of

Doctor of Philosophy

by

Tirthankar Ghosal  
Roll No. 1621CS07

Under the supervision of  
Dr. Asif Ekbal and Prof. Pushpak Bhattacharyya



Department of Computer Science and Engineering  
Indian Institute of Technology Patna  
Patna - 801103, India  
September, 2021

©: 2020 by Tirthankar Ghosal. All rights reserved.



*Dedicated to My Parents.*  
*Mrs. Manoshi Ghosal and Mr. Samaresh Ghosal.*





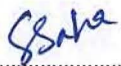
# INDIAN INSTITUTE OF TECHNOLOGY PATNA

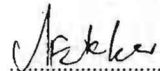
Kanpa Road, Bihta, Patna, Bihar, India

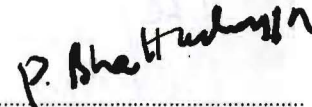
Date: 28.08.2021


## CERTIFICATE OF APPROVAL

Certified that the thesis entitled “STUDIES IN ASPECTS OF PEER REVIEW: NOVELTY, SCOPE, RESEARCH LINEAGE, REVIEW SIGNIFICANCE AND PEER REVIEW OUTCOME” submitted by Mr/Ms. **TIRTHANKAR GHOSAL** (Roll No. 1621CS07) to Indian Institute of Technology Patna, for the award of the degree of Doctor of Philosophy has been accepted by the external examiners and that the student has successfully defended the thesis in the viva-voce examination held today i.e. 28.08.2021.

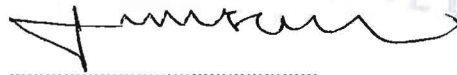
  
CHAIRPERSON of the DC  
(Dr. Sriparna Saha)


  
Supervisor  
(Dr. Asif Ekbal)

  
Supervisor  
(Prof. Pushpak Bhattacharyya)

  
DC Member  
(Dr. Samrat Mondal)

  
DC Member  
(Dr. Sweta Sinha)

  
Internal Examiner  
(Dr. Jimson Mathew)

  
SIGNATURE OF EXTERNAL MEMBER  
(Prof. Ponnurangam Kumaraguru)



## DECLARATION BY THE SCHOLAR

---

I certify that:

- The work contained in this thesis is original and has been done by me under the guidance of my supervisors .
- The work has not been submitted to any other Institute for any degree or diploma.
- I have followed the guidelines provided by the Institute in preparing the thesis.
- I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
- Whenever I have used materials (data, theory and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the reference section.
- The thesis has been checked by anti-plagiarism software.

*Tirthankar Ghosal*

Tirthankar Ghosal





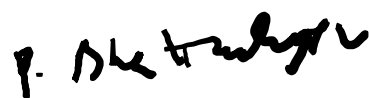
## CERTIFICATE

---

This is to certify that the thesis entitled “**Studies In Aspects of Peer Review: Novelty, Scope, Research Lineage, Review Significance, and Peer Review Outcome**”, submitted by **Mr. Tirthankar Ghosal** to Indian Institute of Technology Patna, is a record of bonafide research work under my supervision and I consider it worthy of consideration for the degree of Doctor of Philosophy of the Institute.



**Dr. Asif Ekbal**  
Supervisor,  
*Associate Professor*, Department of  
Computer Science and Engineering  
Indian Institute of Technology Patna



**Prof. Pushpak Bhattacharyya**  
Co-supervisor,  
*Professor*, Department of  
Computer Science and Engineering  
Indian Institute of Technology Bombay

Place: Indian Institute of Technology Patna  
Date: Monday 6<sup>th</sup> September, 2021



# Acknowledgement

---

I extend my gratitude to

- my supervisors at IIT Patna
- several under-graduate and post-graduate collaborators, interns at IIT Patna
- annotators at AI-NLP-ML Lab, IIT Patna
- members of AI-NLP-ML Lab, IIT Patna and support staff at Department of Computer Science and Engineering
- my mentor and colleagues at the Oak Ridge National Laboratory where I did my internship
- several anonymous reviewers in different conferences and journals
- members of the Doctoral Committee at IIT Patna

for assisting me in their respective capacity to carry out and improvise the research entailed in this thesis. I also thank Visvesvaraya PhD Scheme under Ministry of Electronics and Information Technology (MeitY), Government of India to support me with the fellowship to carry out this research.

Lastly I can never thank my family to believe in me and bear with all my antics in several stages of this thesis.

Place: Indian Institute of Technology Patna  
Date: Monday 6<sup>th</sup> September, 2021

*Tirthankar Ghosal*  
Tirthankar Ghosal



# Abstract

---

The process of peer-review is considered as the *sentinel of science* in scholarly communications. However, with the deluge of research articles and the overload of scholarly information, it is increasingly becoming difficult for humans to keep up with the pace of the latest research. As a result, several problems are crippling the scholarly peer-review process ranging from predatory publishing, incentivized low-quality research, plagiarism, manipulating the review process, timeliness, etc. Here in this work, using Machine Learning (ML) and Natural Language Processing (NLP), we investigate some critical issues relevant to scholarly communications and peer review.

The first problem we investigate as part of this thesis is **Document-level Novelty Detection**. Detecting whether a document contains sufficient new information to be deemed as *novel*, is of immense significance in this age of data duplication. Existing techniques for document-level novelty detection perform mostly at the lexical level and are unable to address the semantic level redundancy. These techniques usually rely on handcrafted features extracted from the documents in a rule-based or traditional machine learning setup. In this thesis, we investigate if we can automatically identify the *newness* in a document based on its information content against a set of relevant documents. We create two benchmark datasets: TAP-DLND 1.0 and TAP-DLND 2.0 for novelty detection at the document level. For TAP-DLND 1.0 we perform annotations at the document level whereas for TAP-DLND 2.0 we perform annotations at the sentence-level. We develop several methods ranging from handcrafted feature engineering to Convolutional Neural Networks (CNN)-based deep neural architectures to attention-based models. Our deep neural model based on feature extraction from a composite *Relative Document Vector* via a deep CNN is the first full-fledged deep architecture for document-level novelty detection. We achieve *state-of-the-art* results on two document-level novelty detection datasets: TAP-DLND 1.0 and APWSJ and on a paraphrase detection dataset Webis-CPC. With our subsequent exploration we develop a decomposable attention-based model which leverages on alignment between source-target document pairs. We achieve *state-of-the-art* accuracies on TAP-DLND 1.0 and APWSJ datasets for the problem. Next we focus on to quantify the amount of *newness*. We encapsulate the source-target information into a semantic unit *Source Encapsulated Target Document Vector (SETDV)* and use a deep CNN to extract meaningful features. We employ the *Two Stage Recall Theory* to select our relevant premises in the process. Our approach SETDV-CNN achieves significance performance on TAP-DLND 2.0 dataset to calculate the *novelty score* of a document. In our final approach towards document novelty, we establish the role of multi-premise entailment towards the task. We use a pre-trained textual entailment model to do *relevance detection* and *novelty detection* in a pipelined two-staged architecture. Our latest method achieved the best results so far on several novelty detection, plagiarism detection datasets. We conduct our experiments for novelty detection on objective newspaper texts due to the lack and difficulty in producing gold-standard data for scientific novelty. We intend to leverage our novelty investigations on news article data for scholarly texts in the future.

The next problem we investigate as part of this thesis is **scope detection in research articles**. One particular challenge in peer review these days is that the journal editors and conference program chairs are overwhelmed with the ever-increasing rise in article submissions. Studies show that many submissions are not well-informed and do not fit within the scope of the intended journal or conference. The relevance of the article is to be ascertained before subjecting it to the formal review process. Here in this work, we investigate how an AI could assist the editors and program chairs in identifying potential *out-of-scope* submissions based on the past accepted papers of the particular journal conference. We experiment with several methods, including feature-based machine learning, deep multimodal architectures, to multiview clustering. Our first method based on hand-crafted feature engineering from several sections of the research article performed better than a popular scholarly article recommender system. The significance of the *bibliography* section towards the problem became clear with our method. With our next exploration we develop a multimodal deep neural network that extracts features from paper full-text and images of the research articles to ascertain its scope. We perform better than several popular baselines and achieve a significant performance in identifying the scope of a research article. Finally, with our third approach we explore the role of *multiview clustering* in the problem with the hypothesis that *outliers* to a cluster of accepted articles could be treated as *out of scope*. Our semi-supervised approach achieved significant performance on a real scholarly dataset comparable to our multimodal deep architecture. Our investigations suggest that we can effectively design automated systems to locate submissions that are not an appropriate fit for the venue’s topical coverage.

Our investigations converged into the problem where we wanted to see how AI would assist in the peer review process. The recent unprecedented growth in paper submissions in major Machine Learning and AI conferences is placing a grand challenge to the community to maintain the high-quality reviewing practices while also ensuring fair evaluation of the manuscripts. We investigate if we can identify which reviews were significant enough to be considered in the decision-making. We design an end-to-end deep network that leverages on *exhaustiveness* and *strength* of the peer review in a multi-task fashion to comprehend its *informativeness*. Since this was a novel problem, there are no comparing systems, however, we perform significantly better than the baselines. Peer review texts contain rich sentiment information of the reviewer reflects his/her overall attitude towards the research conducted and could be a valuable entity to predict the acceptance or rejection of a research work. Further, we study how we could use the sentiment information embedded within peer review texts to help editors/program chairs make better editorial decisions. Our *DeepSentiPeer* model is one of the first attempt to automatically predict decisions in peer review. Our next problem aims to reduce the information overload of the researchers/reviewers. Finding the relevant literature or identifying the works that have inspired the current work under scrutiny is central to evaluation in peer review. In this work, we analyze the context of citations within paper full-text to identify a given research lineage. The idea is to leverage on citation significance classification to build a research lineage via a *significant citation graph*. We argue that finding the lineage of a given research would help identify the true academic impact of a paper beyond quantitative citation counts. In turn, this would help discover significant relevant research to judge the merit of the current submission. Our feature engineering-based approach achieved significant performance improvement over the comparing systems for *citation significance detection*. We demonstrate the idea of forming a *research lineage* with our trained model on two different research topics and hence establish the *proof of concept*.

The current investigation uses scientific methodology to study science itself, which is also better known as Meta-Science or *research on research* or *the science of science*. We document our findings in detail in our papers listed under publications and summarised the corresponding chapters' core experiments and outcomes. Our investigations reveal some interesting observations on each of these problems, which may prove beneficial to incorporate AI solutions in scholarly communications. We also enlist the challenges and issues that one needs to address before enforcing an AI-assisted peer review system. However, there are several other use-cases of AI in peer review, which are not within the present thesis's scope. We seek the community to take the research forward and address the limitation of the problems we investigate in this thesis.

**Keywords:** Peer Review, Scholarly Communications, Natural Language Processing, Machine Learning, Novelty, Scope, Citation Contexts, Review Quality.





# List of Tables

---

3.1	TAP-DLND 2.0 and TAP 1.0 dataset statistics. Inter-rater agreement [1] is measured for 100 documents for sentence-level annotations by two raters. . . . .	27
3.2	Sentence-level annotations. The target document sentences are annotated w.r.t. the information contained in the source documents for each event. The annotations are qualitatively defined. We assign scores to quantify them. . . . .	29
3.3	Feature Set Definition for Novelty Classification [2] . . . . .	30
3.4	Results for Redundant class on APWSJ, <i>LM</i> $\rightarrow$ Language Model, <i>Mistake</i> $\rightarrow$ 100-Accuracy. Except for RDV-CNN, all other numbers are taken from [3] . . . . .	37
3.5	Results for Paraphrase class on Webis-CPC (in %), IDF $\rightarrow$ Inverse Document Frequency, LR $\rightarrow$ Logistic Regression . . . . .	38
3.6	Results on TAP-DLND 1.0, P $\rightarrow$ Precision, R $\rightarrow$ Recall, A $\rightarrow$ Accuracy, R $\rightarrow$ Recall, MLP $\rightarrow$ Multi Layer Perceptron, N $\rightarrow$ Novel, NN $\rightarrow$ Non-Novel, IDF $\rightarrow$ Inverse Document Frequency . . . . .	38
3.7	Results on TAP-DLND 1.0, P $\rightarrow$ Precision, R $\rightarrow$ Recall, A $\rightarrow$ Accuracy, $F_1$ $\rightarrow$ Average F-Score, N $\rightarrow$ Novel, NN $\rightarrow$ Non-Novel, * $\rightarrow$ measures from [3] with Logistic Regression (LR), ‡ $\rightarrow$ measure from [4] with Logistic Regression (LR), † $\rightarrow$ 10-fold cross-validation output, IDF $\rightarrow$ Inverse Document Frequency . . . . .	50
3.8	Results for Redundant/Non-Novel (NN) class on APWSJ, <i>LM</i> $\rightarrow$ Language Model, <i>Mistake</i> $\rightarrow$ 100-Accuracy, † $\rightarrow$ 10-fold cross-validation output, * $\rightarrow$ results taken from [3] . . . . .	51
3.9	Performance of the proposed approach against the baselines and comparing systems, PC $\rightarrow$ Pearson Correlation Coefficient, MAE $\rightarrow$ Mean Absolute Error, RMSE $\rightarrow$ Root Mean-Squared Error, Cosine $\rightarrow$ Cosine similarity between predicted and actual score vectors . . . . .	64
3.10	Sample text from Webis-CPC-11 to simulate the high-level semantic paraphrasing in the dataset. . . . .	77
3.11	Sample from P4PIN to show plagiarism (non-novel) instance . . . . .	78
3.12	Sample from Wikipedia Rewrite Dataset to show a plagiarism (non-novel) instance . . . . .	78
3.13	Results on TAP-DLND 1.0, P $\rightarrow$ Precision, R $\rightarrow$ Recall, A $\rightarrow$ Accuracy, R $\rightarrow$ Recall, N $\rightarrow$ Novel, NN $\rightarrow$ Non-Novel, 10-fold cross-validation output . . . . .	81
3.14	Results for redundant class on APWSJ, <i>Mistake</i> $\rightarrow$ 100-Accuracy. Except for [3] all other figures correspond to a 10-fold cross-validation output . . . . .	82
3.15	Performance of the proposed approach against the baselines and comparing systems TAP-DLND 1.1, PC $\rightarrow$ Pearson Correlation Coefficient, MAE $\rightarrow$ Mean Absolute Error, RMSE $\rightarrow$ Root Mean-Squared Error, Cosine $\rightarrow$ Cosine similarity between predicted and actual score vectors . . . . .	82
3.16	Results for paraphrase class on Webis-CPC, 10-fold cross-validation output . . . . .	83

4.1	Scope-Check figures for <i>out-of-scope</i> (OS) class across 6 journals, $P \rightarrow Precision$ , $R \rightarrow Recall$ , $\ddagger \rightarrow$ Baseline using only Title and Abstract with SVM classifier. The Accuracy values ( $\dagger$ ) for <i>ScopeJr</i> are statistically significant over EJF performance (two-tailed t-test, $p < 0.05$ ) . . . . .	103
4.2	Dataset-I Statistics (Elsevier), FT $\rightarrow$ Full-Text, Actual Negative are the instances (papers) which were desk-rejected due to <i>out-of-scope</i> from the concerned journal, Bib $\rightarrow$ Bibliography, J1 $\rightarrow$ ARTINT, J2 $\rightarrow$ COMNET, J3 $\rightarrow$ STATPRO, J4 $\rightarrow$ JNCA, J5 $\rightarrow$ SIMPAT, J6 $\rightarrow$ CSI, Tr $\rightarrow$ Training Data, Tt $\rightarrow$ Test Data . . . . .	107
4.3	Dataset-II Statistics (Open Access AI/ML/NLP/CV Papers), This statistics signify the volume of information processing corresponding to the three modalities .	108
4.4	Scope Detection (Binary Classification) Results on Dataset-I (Elsevier Journals), Accuracy in % . . . . .	115
4.5	Results on Dataset-II (AI/ML/NLP/CV). Multi-class classification. . . . .	117
4.6	Cluster Prediction ( <i>In-Scope</i> or <i>Out-Scope</i> ) Results on the 3 journals, $\dagger \rightarrow$ Baselines	121
5.1	Dataset Statistics . . . . .	127
5.2	Results on Aspect Score Prediction Task. Training is done with only ICLR 2017 papers/reviews, $\dagger \rightarrow$ Cross-Domain: Training on ICLR and testing upon entire data of ACL/CoNLL available in PeerRead dataset, $\ddagger \rightarrow$ Test set is kept the same as [5], RMSE $\rightarrow$ Root Mean Squared Error. CNN variant as in [5] is used as the comparing system. . . . .	132
5.3	Results on Accept/Reject Classification Tasks. Training is done with ICLR 2017+ICLR 2018 papers/reviews, $\dagger \rightarrow$ Cross-Domain: Training on ICLR and testing upon the entire data of ACL/CoNLL, $\ddagger$ Test Set is kept the same as [5], RMSE $\rightarrow$ Root Mean Squared Error, $*$ $\rightarrow$ 65.79% if only trained with ICLR 2017, Comparing System [5] is feature-based and considers only paper, and not the reviews. . . . .	132
5.4	Pearson Correlation (PC) Coefficient between the <i>Recommendation Scores</i> and <i>Sentiment Activations</i> . This is to account for the fact that sentiment is actually correlated with the prediction signifying the strength of the model. . . . .	134
5.5	A qualitative study of the effect of sentiment in the overall recommendation score prediction. Prediction $\rightarrow$ is the overall recommendation score predicted by our system, Actual $\rightarrow$ is the recommendation score given by reviewers. <b>Senti_Act</b> are the output activations from the final layer of <i>MLP_Senti</i> which are augmented to the decision layer for final recommendation score prediction. The correspondence between the sentiment embedded within the review texts and Sentiment Activations are fairly visible in Figure 5.3. Kindly refer to Figure 5.5 for polarity strengths in individual review sentences. The OpenReview links in the table above give the full review texts. . . . .	134
5.6	Example reviews illustrating <i>Exhaustiveness</i> and the <i>Intensity</i> components of our definition of the <i>Quality</i> of an academic peer-review. Note that these two components manifest differently in different reviews. . . . .	143
5.7	Dataset Statistics . . . . .	146
5.8	Performance of Score Prediction(Regression Task) across all the models. Layer1, Layer2, and Layer3 denotes <i>Exhaustive</i> , <i>Reviewer aspect</i> , <i>Intensity</i> tasks respectively. Training is done on ICLR 2017, 2018 and 2019 data, and testing is done on held out test data. RMSE $\rightarrow$ Root Mean Square Error. . . . .	153
5.9	Review Significance Score Predictions . . . . .	154
5.10	Results on Citation Significance Detection on Valenzuela dataset . . . . .	165
5.11	Classification Result of various Classifiers for Citation Significance . . . . .	165

## LIST OF TABLES

---

6.1	Summarizing contribution of this thesis . . . . .	174
-----	---	-----



# List of Figures

---

3.1	The TAP-DLND 1.0 corpus structure. We retain the structure in the extended dataset we use in the current work . . . . .	26
3.2	The Sentence-Level Annotation Interface used to generate the Document-Level Novelty Score (Gold Standard) . . . . .	28
3.3	Significance of features <i>based on Information Gain (IG)</i> . The length of the bar corresponds to the average merit ( $X : IG$ ) of the feature ( $: Y$ ). . . . .	31
3.4	RDV-CNN framework for Novelty Detection. Generic SNLI Training [6]. The sentence encoder is trained on SNLI. The RDV-CNN is trained with the respective novelty datasets. . . . .	33
3.5	Overall architecture for document-level novelty detection. Component (b) is the inner-attention sentence encoder . Component (c) shows how the inner attention sentence encoder is trained on the SNLI corpus. Component (a) is the sentence-level decomposable attention model we use in our work for document level novelty detection. $s_{11}$ , $s_{12}$ represents the two sentences in source document $d_1$ in the example introduced in Section 3.6. $s_{21}$ , $s_{22}$ are the two sentences in source document $d_2$ . $d_1$ and $d_2$ are concatenated to form a single source document. $t_{21}$ , $t_{22}$ are the two sentences in target document $d_4$ . Simply reading the example we can conclude that $t_{21}$ and $t_{22}$ directly follow from $s_{11}$ and $s_{22}$ . $d_4$ is redundant if we consider $d_1$ and $d_2$ as the source documents. . . . .	44
3.6	Attention matrix visualization via heat map for the example in Section 3.6. $d1$ and $d2$ are concatenated to form the source document. $d4$ is the target document. . . . .	53
3.7	Attention matrix visualization via heat map for a correctly predicted (a) novel and (b) non-novel document from TAP-DLND 1.0. (a) Many dark patches signify that most of the target sentences are not highly attending to any source sentence. Hence, may contain sufficient new information. (b) Lesser dark patches indicate that the target sentences are highly aligned to the source sentences and may contain redundant information. Wrongly predicted instances $\rightarrow$ (c), (d), (e). . . . .	56
3.8	The overall Novelty Score Prediction Architecture (SETDV-CNN) with 12 sentences in the target document (T1) and S1, S2, S3 $\rightarrow$ source documents . . . . .	59
3.9	Scatter plot of Actual (Gold Standard) vs Predicted (Proposed) Document-Level Novelty Score . . . . .	65
3.10	Multi-Premise Entailment Based Document-Level Novelty Detection Architecture . . . . .	71
3.11	Predicted novelty scores for documents in P4PIN and WikiRewrite by our model trained on TAP-DLND 1.0 . . . . .	84

3.12	Heatmap depicting the attention scores between the source and target document (Example 2 in section 3.8.2). $t_1, t_2$ are the target document sentences (vertical axes) and $s_1, s_2, s_3, s_4$ are source document sentences (horizontal axes). The brighter the shade, more is the alignment, signifying an affinity towards non-novelty	86
3.13	Heatmap depicting the attention scores between the source ( $S_1, S_2, S_3$ ) and target document ( $T_1, T_2$ ). The brighter the shade, more is the alignment, signifying an affinity towards non-novelty	87
3.14	Heatmap of the misclassification instance	88
4.1	Box plots of various factors across an exclusive set of 1000 IS and 1000 OS articles. The match is in terms of overlap of keywords, referenced paper titles, bibliographic venues and authors with respect to past accepted papers of each journal. The median is always high for IS w.r.t. OS articles.	98
4.2	Significance of features observed by ranking features based on Information Gain	104
4.3	Proposed Deep Multimodal Neural Architecture for Scope Detection	109
4.4	Multi-view Clustering of <i>In-Scope</i> and <i>Out-of-Scope</i> articles, X-axis→Semantic View, Y-axis→Lexical View, Z-axis→Bibliography View	119
5.1	Pearson Correlation of Review Sentiment (:X) with different Aspect Scores (:Y) on ACL 2017 dataset. A1→Appropriateness, A2→Clarity, A3→Impact, A4→Meaningful Comparison, A5→Originality, A6→Recommendation, A7→Soundness/Correctness, A8→Substance, D→Decision. pos→Positive Sentiment Score, neg→Negative Sentiment Score, neu→Neutral Sentiment Score, com→Compound Sentiment Score. To calculate the sentiment polarity of a review text, we take the average of the sentence wise sentiment scores from Valence Aware Dictionary and sEntiment Reasoner (VADER) [7].	127
5.2	<i>DeepSentiPeer</i> : A Sentiment Aware Deep Neural Architecture to Predict Reviewer Recommendation Score. Decision-Level Fusion and Feature-Level Fusion of Sentiment are shown for Task 1 and Task 2, respectively.	128
5.3	Projections of the output activations of the final layer of <i>MLP_Senti</i> . Points are annotated for Reviews from Table 5.5. X: Predicted Recommendation Scores, Y: Sentiment Activations	134
5.4	Normalized Confusion Matrix for Accept/Reject Decisions on ICLR 2017 test data with <i>DeepSentiPeer</i> (Paper+Review+Sentiment) model.	137
5.5	Heatmaps of the sentence-wise VADER sentiment polarity of reviews considered in Table 5.5. Reviews generally reflect the polarity of the reviewer towards the respective work. $s_0 \dots s_n \rightarrow$ are the sentences in the peer review texts.	138
5.6	a) Relative importance of labels for Review Exhaustiveness (Layer 1). b) Distribution of the <i>Exhaustiveness</i> scores for the annotated reviews. c) Distribution of the <i>Reviewer Aspect</i> scores for the annotated reviews (Layer 2)	145
5.7	Our architecture for predicting the significance of peer reviews (main task). The model is trained on the peer review decisions (secondary task)	148
5.8	The cross-attention and sequential coder module from Figure 5.7	149
5.9	Research Lineage	160
5.10	Feature importance ranked via Information Gain. Number of words between citances (F16) and Number of normalized citations (F7) are the most contributing features.	166

## LIST OF FIGURES

---

5.11	Significant Citation graph for a set of papers on <i>Document-Level Novelty Detection</i> . Please refer to the bibliography for the paper details. P1→[8], P2→[9], P3→[10], P4→[11], P6→[12], P7→[13], P8→[14], P9→[15], P11→[16], P12→[17], P17→[18], P22→[19], P23→[20], P24→[4], P25→[3], P28→[21] . . . . .	168
5.12	Significant Citation graph for a set of papers on <i>MENNDL HPC algorithm</i> . Please refer to the bibliography for the corresponding paper details. P1→[22], P4→[23], P6→[24], P7→[25], P8→[26], P9→[27], P12→[28], P13→[29], P14→[30], P15→[31], P16→[32], P17→[33], P18→[34], P19→[35], P20→[36], P25→[3], P28→[21] . . .	169
7.1	TAP-DLND 2.0 Annotation Interface . . . . .	182





# List of Notations and Operations

---

where otherwise explicitly specified the following symbols would mean:

$S_i$	source sentence $i$
$h_t$	hidden state at timestep $t$
$p$	premise
$h$	hypothesis
$b$	bias in neural network
$f$	a non-linear function
$c$	feature-map
$W$	weight-matrix
$\alpha_i$	attention values
$F/G$	feed-forward network
$y$	classification output
$\cos$	cosine similarity
$Pr$	probability
$f_{mlp}$	multi layer perceptron
$\theta_{mlp}$	parameters of a mlp
$\mathbb{R}$	real number space
$\oplus$	vector concatenation operator
$\overrightarrow{LSTM}$	forward pass of the LSTM
$\overleftarrow{LSTM}$	backward pass of the LSTM
$  $	concatenation
$\sigma$	observation noise parameter of the model



# Abbreviations

---

NLP	Natural Language Processing
ML	Machine Learning
AI	Artificial Intelligence
IR	Information Retrieval
DLND	Document-Level Novelty Detection
TDT	Topic Detection and Tracking experiments
TREC	Text Retrieval Conferences
RTE	Recognizing Textual Entailment
APWSJ	Associated Press-Wall Street Journal corpus
GPT-3	Generative Pre-trained Transformer 3
SEO	Search Engine Optimization
IG	Information Gain
AMOS	Archived Multi-objective Simulated Annealing
CNN	Convolutional Neural Network
RDV	Relative Document Vector
LSTM	Long Short Term Memory
SNLI	Stanford Natural Language Inference dataset
SVM	Support Vector Machine
ANN	Artificial Neural Network
RSV	Relative Sentence Vector
SGD	Stochastic Gradient Descent
RF	Random Forest
LR	Logistic Regression
OOV	Out Of Vocabulary words
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network
TAC	Text Analytics Conference
ReLU	Rectified Linear Unit
TE	Textual Entailment
NLI	Natural Language Inference
SENT_DIM	Sentence Dimension
VADER	Valence Aware Dictionary for Sentiment Reasoning
RNN	Recurrent Neural Network
Bi-LSTM	Bi-directional LSTM
SETDV	Source Encapsulated Target Document Vector
tf-idf	term frequency-inverse document frequency
ESIM	Enhanced Sequential Inference Model

NS	Novelty Score
PC	Pearsson Co-efficient
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
KLD	Kullback Leibler Divergence
SAT	Source Aware Target representation
IS	In-Scope articles
OS	Out of Scope articles
DR-OOS	Desk Rejected due to Out of Scope
CV	Computer Vision
CitE	Citation Effect
EJF	Elsevier Journal Finder
KWScore	Keyword Match Score
ADPF	Author Domain Publication Frequency
BoW	Bag of Words
VGG-16	a CNN from Visual Geometry Group, Oxford
RAKE	Rapid Automatic Keyword Extractor
YAKE	Yet Another Keyword Extractor
ICLR	International Conference on Learning Representations
HPC	High Performance Computing
MENNDL	Multi-node Evolutionary Neural Networks for Deep Learning
kNN	k-Nearest Neighbour classification algorithm
SVM	Support Vector Machines

# Contents

---

---

<b>Certificate of Approval</b>	<b>v</b>
<b>Declaration</b>	<b>vii</b>
<b>Certificate</b>	<b>ix</b>
<b>Acknowledgement</b>	<b>xi</b>
<b>Abstract</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xvi</b>
<b>List of Figures</b>	<b>xxi</b>
<b>List of Symbols</b>	<b>xxv</b>
<b>List of Abbreviations</b>	<b>xxvii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction and Motivation . . . . .	2
1.2 Roadmap and Contributions . . . . .	5
1.2.1 Textual Novelty Detection . . . . .	5
1.2.2 Scope Detection . . . . .	6
1.2.3 Investigating AI for Peer Review and Establishing a Research Lineage . .	7
1.3 Chapter Summary . . . . .	9
<b>2 Literature Review</b>	<b>11</b>
2.1 Literature Review on Textual Novelty Detection . . . . .	12
2.1.1 Sentence-level Novelty Detection . . . . .	12
2.1.2 Document-level Novelty Detection . . . . .	13
2.1.3 Diversity Detection . . . . .	14
2.2 Literature Review on Scope Detection . . . . .	14
2.2.1 Academic Recommender Systems . . . . .	15
2.3 Literature on AI for Peer Review . . . . .	16
2.3.1 Peer Review Quality . . . . .	17
2.3.2 Literature Review on Finding a Research Lineage . . . . .	19
2.4 Chapter Summary . . . . .	20

<b>3</b>	<b>Textual Novelty Detection</b>	<b>21</b>
3.1	Introduction . . . . .	22
3.1.1	Why Textual Novelty Detection is important? . . . . .	22
3.1.2	Our Exploration on Textual Novelty Detection . . . . .	23
3.2	TAP-DLND 1.0 corpus . . . . .	24
3.3	TAP-DLND 2.0 corpus . . . . .	27
3.4	Feature-Engineering Approach for Novelty Detection . . . . .	30
3.5	Relative Document Vector-Convolutional Neural Network . . . . .	32
3.5.1	Proposed Method . . . . .	32
3.5.2	Results and Discussion . . . . .	36
3.5.3	Observations and Analysis . . . . .	39
3.6	Decomposable Attention for Novelty Detection . . . . .	41
3.6.1	Proposed Method . . . . .	42
3.6.2	Evaluation . . . . .	48
3.6.3	Results and Discussions . . . . .	50
3.6.4	Analysis . . . . .	53
3.6.5	Error Analysis . . . . .	54
3.7	Quantifying Novelty . . . . .	57
3.7.1	Problem Definition . . . . .	57
3.7.2	Methodology . . . . .	58
3.7.3	Experiments and Results . . . . .	62
3.7.4	Results and Discussion . . . . .	64
3.7.5	Error Analysis . . . . .	65
3.8	Leveraging Multi-Premise Textual Entailment for Novelty Detection . . . . .	67
3.8.1	Textual Novelty Detection: An Entailment Perspective . . . . .	67
3.8.2	Multi-premise Entailment for Novelty Detection . . . . .	69
3.8.3	Encompassing Multiple Premises for Document-Level Novelty Detection . . . . .	70
3.8.4	Datasets for Allied Tasks . . . . .	76
3.8.5	Evaluation . . . . .	78
3.8.6	Results . . . . .	80
3.8.7	Results on Related Tasks . . . . .	83
3.8.8	Analysis . . . . .	85
3.8.9	Error Analysis . . . . .	87
3.8.10	Insights gained from the work . . . . .	89
3.8.11	Limitations of our research . . . . .	89
3.8.12	Transcending to Scientific Novelty - How difficult is the problem? . . . . .	89
3.8.13	Conclusion . . . . .	90
3.9	Chapter Summary . . . . .	91
<b>4</b>	<b>Scope Detection of Research Articles</b>	<b>93</b>
4.1	Introduction . . . . .	94
4.2	Problem Definition . . . . .	95
4.3	Scope Detection . . . . .	95
4.3.1	Desk-Rejection Observations . . . . .	96
4.3.2	Scope of a Journal . . . . .	96
4.4	Feature-based Machine Learning for Scope Detection . . . . .	97
4.4.1	Data Description and Preprocessing . . . . .	97
4.4.2	Methodology . . . . .	99
4.4.3	Evaluation . . . . .	101

## CONTENTS

---

4.4.4	Results and Observations . . . . .	103
4.5	Multimodal Scope Detection . . . . .	105
4.5.1	Problem Definition . . . . .	105
4.5.2	Data Description and Analysis . . . . .	106
4.5.3	Methodology . . . . .	109
4.5.4	Experimental Setup . . . . .	113
4.5.5	Results and Analysis . . . . .	114
4.5.6	Conclusions . . . . .	117
4.6	Multiview Clustering for Scope Detection of Scientific Articles . . . . .	118
4.6.1	Dataset Description/Preprocessing . . . . .	119
4.6.2	Methodology . . . . .	119
4.6.3	Evaluation . . . . .	120
4.6.4	Conclusion . . . . .	121
4.7	Insights gained from this work . . . . .	121
4.8	Limitations . . . . .	121
4.9	Chapter Summary . . . . .	122
<b>5</b>	<b>AI in Peer Review and Finding a Research Lineage</b>	<b>123</b>
5.1	Introduction . . . . .	124
5.2	Predicting Peer Review Outcome . . . . .	124
5.2.1	Problem Definition . . . . .	126
5.2.2	Data Description and Analysis . . . . .	126
5.2.3	Methodology . . . . .	127
5.2.4	Experimental Setup . . . . .	131
5.2.5	Results and Analysis . . . . .	131
5.2.6	Conclusion . . . . .	139
5.3	Peer Review Significance . . . . .	140
5.3.1	Quality of a Review . . . . .	142
5.3.2	Dataset Description . . . . .	146
5.3.3	Methodology . . . . .	147
5.3.4	Evaluation . . . . .	151
5.3.5	Results . . . . .	152
5.3.6	Qualitative Analysis . . . . .	153
5.3.7	Ethical Issues . . . . .	156
5.3.8	Conclusion . . . . .	157
5.4	Finding a Research Lineage . . . . .	158
5.4.1	Research Lineage . . . . .	160
5.4.2	Dataset Description . . . . .	161
5.4.3	Methodology . . . . .	161
5.4.4	Evaluation . . . . .	164
5.4.5	Conclusion and Future Work . . . . .	169
5.5	Limitations of our work . . . . .	170
5.6	Chapter Summary . . . . .	171
<b>6</b>	<b>Conclusions and Future Works</b>	<b>173</b>
6.1	Summary . . . . .	173
6.2	Contributions of the Thesis . . . . .	174
6.3	Limitations of the Thesis . . . . .	175
6.4	Conclusions . . . . .	175

6.5 Future Work . . . . .	176
<b>7 Appendix</b>	<b>179</b>
<b>References</b>	<b>183</b>
<b>List of Publications</b>	<b>197</b>



# CHAPTER 1

---

## Introduction

---

We begin this chapter with a high level description of the thesis, *Studies In Aspects of Peer Review: Novelty, Scope, Research Lineage, Review Significance, and Peer Review Outcome*, explain the primary motivation behind the proposed research work and its importance in solving some critical problems related to academic peer-review. In this chapter, we also briefly introduce the problems and highlight our contributions. Finally we conclude by sketching the scope of the overall research work and provide a brief outline for each of the subsequent chapters.

---

## 1.1 Introduction and Motivation

Peer Review is at the heart of scholarly communications and the cornerstone of research validation. The manifestation of human scientific progress is predominantly in the form of research papers. However, with the deluge of research articles leading to the overload of scholarly information [37], it is increasingly becoming difficult for humans to keep up with the pace of the latest research. A report back in 2012 by the International Association of Scientific, Medical, and Technical Publishers<sup>1</sup> indicates that in mid-2012, there were 28,100 active scholarly peer-reviewed journals, and the number of articles, journals, and researchers continues to grow steadily. One can hardly imagine the volume of scientific literature and its accelerated pace of growth now. As a result, the scholarly peer-review process is suffering from several problems including predatory publishing [38], incentivized low-quality research [39], plagiarism [40], gaming the system [41], timeliness [42], etc. Taming this huge volume of scientific information overload is a frontier for the Natural Language Processing (NLP) and Machine Learning (ML) community.

Peer review is the process that ensures rigorous review and validation of research results and methodology. A journal editor sends a researcher’s manuscript to a team of experts that examines the research methodology, results, analysis, and conclusions to determine and ensure material accuracy, relevance, and importance — and decide whether it is a good fit for the publication. But peer review has its issues [43]. Some scientists advocate abandoning the process [44], but most believe it is a necessary practice that could be improved [45]. Some researchers are using scientific methodology to study science itself is also better known as Meta-Science or *research on research* or *the science of science* [46].

The peer-review process has several inherent challenges, from the potential bias among peer reviewers, to process integrity of publications to the length of time to publication. Many researchers and publishers suggest that AI could eliminate or reduce some of these challenges [47, 48]. Some areas in the scientific publishing life-cycle where AI could make an impact are:

- *Peer Reviewer Selection:* Journal editors typically select half a dozen potential reviewers based on academic credentials, fields of expertise, previous reviews, possible conflicts of interest, workloads, and any other relevant criteria the editor desires. AI can help find potential new reviewers from online sources and do the paper-reviewer pairing from a pool [49, 50].
- *Peer reviewer bias:* Sometimes reviewers may trade favorable reviews for one another’s works, or others may have personal prejudices for or against specific authors. One can

---

<sup>1</sup>[https://www.stm-assoc.org/2012\\_12\\_11\\_STM\\_Report\\_2012.pdf](https://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf)

## 1.1 Introduction and Motivation

---

employ AI systems to address the issue of reviewer bias [51] by using specific criteria to select the reviewers and screen them for any bias towards a particular author or topic. The open peer review model [52] aims to mitigate this bias via publishing the actual review and the reviewer's name.

- *Pre-screening:* AI algorithms can be used to check for plagiarism [53], author verification [54], impact factor prediction [55], proper methodology, fake data, data gaps, and faulty analysis and conclusions. The algorithm might also look at consistency through the paper; for example, searching for statistical error or method description incompleteness: if there is a multiple group comparison, whether the p-value correction algorithm is indicated. Statcheck [56] is such an attempt towards this direction.
- *General automation:* In addition to improving the reviewer selection process and data checking, AI can help with general tasks such as maintaining databases of authors, manuscripts, and reviewers; maintaining workflows such as sending reviewer-author communications, notifying authors of paper status, sending thank you notes, selecting alternate reviewers and re-sending manuscripts; and other basic record-keeping tasks. Elsevier used to employ a system called EVISE<sup>2</sup> to perform some of these tasks.
- *Time to publish:* The scientific publication process is time-consuming, and there could be many reasons for delays [57]. When an author submits a scientific manuscript to when it is finally published, it can be months, even years. The back-and-forth between reviewers and author and the resulting corrections and modifications are time-consuming, not to mention the ever-increasing number of papers. By streamlining the reviewer selection process, pre-screening research papers for objectivity, improper methodology, incorrect or insufficient results, and automating many record-keeping tasks, the time to publication can be reduced to days or weeks, which speeds up subsequent research based on scientific findings.
- One can use Machine Learning to predict the future impact of a given work (e.g., future citation counts [58]), and in effect to do the job of impact analysis and assist in decision-making alongside a human reviewer.

There are also some criticisms of AI in peer review. In a study [59], authors present fake peer reviews to academicians, who were then asked to either agree or disagree with the outcomes—one-quarter agreed with them. Analysts raise questions on the impact of such systems on the practice of scholarly writing, such as how authors may change their approach when they know a machine will evaluate their manuscript<sup>3</sup>, or how machine assessment could discover unfounded authority

---

<sup>2</sup><https://www.evise.com>

<sup>3</sup><https://www.wired.com/2017/01/peer-review-shortcomings-ai-risky-fix/>

in statements by authors through analysis of citation networks [60]. One additional potential drawback of automation of this sort is the possibility of detecting false positives that might discourage authors from submitting.

Some platforms already incorporate such AI-assisted methods for a variety of purposes. Scholastica<sup>4</sup> includes real-time journal performance analytics to assess and improve the peer-review process. The Journal of High Energy Physics uses automatic assignment to editors based on a keyword-driven algorithm [61]. Initiatives such as Meta<sup>5</sup>, an AI tool that searches scientific papers to predict the trajectory of research, highlight the great promise of artificial intelligence in research and for application to peer review.

As the larger AI community is embarking onto some difficult problems like mastering intellectual games [62], predicting protein-structure folds [63], we wonder is it the time to tackle some formidable challenges in science like to determine *what is new knowledge*, which is so critical in scientific research? Can it be done via text mining? Can the review process be automated to some extent to flag inappropriate, *out-of-scope* manuscripts to assist the editors and thereby speed-up the peer review process?

We commenced the investigation to objectively study the peer review process back in 2016. We intended to uncover the potential use-cases where NLP and ML could play an assistive role. We started with the study of rejected paper reviews from 17 different computer science journals. Our survey of about 7000+ reviews revealed that the significant causes of concern of the reviewers are: *Novelty* of the work, whether the article was in *Scope*, and if the *Quality* of the work stands up to the standards of the journal. All these criteria are standard yet are very abstract and not straightforward to define. We began our investigations to explore these problems under the lens of NLP and ML.

The major challenge we encountered while working with *textual novelty* is the lack of gold-standard data for the document-level variant of the problem. Defining scientific novelty is not straightforward, and producing ground-truth data for scientific novelty would require high-level domain expertise. Hence, we began our investigation to study textual novelty with more objective newspaper data. We detail the corresponding data preparation in Chapter 3.

For the other problems (entailed in subsequent sections), we carry our investigation with scholarly texts from different journals and conferences.

Later in the journey, we became interested in exploring how NLP/ML models can analyze the peer review texts to predict peer review outcomes and judge the significance of the peer reviews.

---

<sup>4</sup><http://www.scholasticahq.com>

<sup>5</sup><http://www.meta.com>

## 1.2 Roadmap and Contributions

The motivation behind this thesis is to explore some uncharted frontiers where AI/ML can play a role in scholarly communications to counter the huge information overload. Specifically, we are probing the following problems:

- *Document-level Novelty Detection* with general text.
- *Scope Detection* of research articles.
- Analyzing peer-review texts for their *significance* in decision-making and predicting the peer review *outcome*.
- Analyzing context of citations within paper full-text to trace the *lineage* of given research, thus also identifying the impact or *quality* of the work.

The current chapter provides the context and motivation of this thesis. In chapter 2, we discuss the relevant literature about the different problems we investigate.

### 1.2.1 Textual Novelty Detection

The most significant aspect of judging an article’s merit in scholarly communications is its *Novelty*. Literature reveals that there has been very less work on *Document-Level Novelty Detection*. However, we can trace a good number of investigations on sentence-level novelty detection that forms the base of our research into the problem.

Document-level novelty detection implies categorizing a document as novel if it contains sufficient new information with respect to whatever relevant is previously known or seen. Until now, document-level novelty detection was primarily seen as an Information Retrieval (IR) problem with a focus on retrieving sentences that consist of new information from a given candidate set. With our foray, we went on to explore if we could identify an entire document as being novel, non-novel, or partially-novel based on its information content against a set of source documents already scanned/seen by the system. Sentence-level novelty detection, although important but has comparatively fewer use-cases than its document-level counterpart. The first challenge we faced while exploring the concerned task was the lack of document-level annotated data. So our point of departure was to create datasets for our experiments.

Chapter 3 details our explorations on the problem. Novelty annotations for scholarly data are not straightforward and costly to produce. Also, defining scientific novelty in terms of only text is not enough. Hence we carry on our experiments with more objective newspaper data. The

idea is to leverage the knowledge gained from experiments on novelty detection on newspaper texts to scientific texts in the future. We sum up our **contributions** to the problem:

- Developed a document-level novelty detection dataset: TAP-DLND 1.0 [2]
- Developed a sentence-level annotated dataset for document-level novelty scoring: TAP-DLND 2.0 (under communication)
- Devised a simple feature-engineering approach with traditional machine-learning algorithms for novelty classification [2]
- A deep-neural architecture leveraging using a convolutional neural network [64]
- A deep-neural approach for quantifying the amount of new information in a document [65]
- A decomposable-attention based model for document-level novelty detection [66].
- Explored multi-premise textual entailment for textual novelty detection (under communication)

### 1.2.2 Scope Detection

The very first stage in peer review is the initial screening, usually performed by editors. Day by day, the number of submissions made to each journal is rising. Editors, who are usually full-time academicians, are overwhelmed with these huge number of manuscripts they had to go manually through. With the ever-expanding volume of research articles, it is increasingly becoming difficult for the editors to keep pace with the latest research and trends. These also hinders editorial response in a reasonable time frame. At this stage, journal editors are entrusted to take either of the two decisions: whether to forward the manuscript to expert reviewers for meticulous evaluation or to outright reject the paper from the desk.

We did extensive feature engineering on the reviews of rejected papers (7000+ desk-rejected reviews from 17 different Computer Science journals). We found that in spite of having merit, around 25-30% [67] of the articles are rejected from the editor’s desk simply because they do not fall into the ”scope” of the journal. Those papers are rejected from the editor’s desk, and authors made less-informed decisions while submitting their manuscripts. But again, having a strict definition of the ”Scope” of a journal is not straightforward since different journals have different views of scope (review papers, technical papers, open domain, etc.). However, we take a simplistic assumption that earlier accepted papers of a given journal may indicate its topical coverage and domain of operation. We went on to explore: can a machine trained on earlier accepted papers also predict if an incoming paper falls within the domain of operation or scope

## 1.2 Roadmap and Contributions

---

of the journal concerned? We went on to view this task as a document classification problem in machine learning and devise our methods.

Chapter 4 of this thesis details our explorations for *Scope Detection* of academic manuscripts. Our contributions in this chapter are:

- Proposed a feature-based machine learning method to identify out-of-scope submissions [68, 69].
- Proposed a multimodal deep neural architecture to classify article submissions based on their aptness to the concerned venue [70].
- Proposed a multiview clustering algorithm to identify publications outside the cluster of accepted articles for a given venue [71].

### 1.2.3 Investigating AI for Peer Review and Establishing a Research Lineage

In Chapter 5, we explore three important problems related to scholarly communications and peer review. The objective is to see how AI can assist in peer review, and predict peer review outcome, judge the significance of a peer review. We also explore if we can leverage on identifying significant citations to establish a lineage of a given research.

#### Predicting Peer Review Outcome

With the first problem, we investigate if we can develop an AI trained on earlier papers of a venue and predict the peer review outcome [72]. We explore how reviewers' sentiment encoded within peer review texts can predict the final decision and if we could further leverage it for AI-assisted decision-making.

In this work, we investigate a novel problem to predict the outcome of a peer-review process by extracting features from papers and peer review texts via a deep neural network. We learn representations of the paper full-text and peer-review texts, extract sentiment of the reviewers embedded within the peer review text, compose these learned representations, and see if our AI could predict the final decision. The motivation here is to present the AI as the fourth reviewer to provide an additional layer of confidence to the editors/area chairs. The deep model learns the high-level interplay between the paper and the reviews and the corresponding sentiment from the earlier accepted and rejected papers and predicts the decision for a new paper. We achieve encouraging results in this experiment and learn that we may need to justifiably grade the reviews before employing an AI for this crucial task. This insight paved our way for the subsequent problem. We crawl and prepare our dataset from the OpenReview<sup>6</sup> platform, which

---

<sup>6</sup><https://openreview.net>

hosts the peer review data for accepted and rejected papers.

### Significance of Peer Reviews

With the second problem in this chapter, we try to decipher whether a given peer review was significant or not.

We could hardly find a researcher who is fully satisfied with the reviews he/she has received in a rejection decision. With the ever-increasing peer review workload, the quality of peer-reviews is suffering. Some reviews are detailed, exhaustive, and constructive, while some are trivial and done in haste with little take-aways. We undertake this new task, *can we isolate trivial reviews so that they have less impact on the decision-making? Can we identify significant peer reviews and also grade the reviews so that our AI could predict the final decision by assigning justifiable weights to the reviews?* We understand that we need high-quality annotated peer review data for this crucial task, which we prepared in the process. For this work, we assume that the editors/area chairs have implicitly placed due importance on significant reviews while arriving at their decisions. In this work, since we do not have annotated data for review significance, we assume decision prediction to the scaffolding task to the main task of review significance prediction. We define review significance detection as a function of review exhaustiveness and embedded review sentiment. We formulate review exhaustiveness as a neural cross-attention mechanism between the paper and review representation. In a typical use case, the editor can have the decision of the AI (as the fourth reviewer) on the paper with significance scores on each of the human reviews.

This investigation deals with analyzing peer review texts and investigating whether the peer review was helpful to the authors, and grade peer reviews based on their perceived significance.

### Establishing a Research Lineage via Significant Citations

With the third problem, we explore if we can trace a research lineage via identifying significant citations. The idea is to determine how one research stems from another, and if that knowledge propagation is visible in a scholarly network.

As we all know that the current measure of research quality is citation counts and h-indices, and these measures incentivize the authors, the concerned research, and also the venue where published. These measures are quantitative and put equal weights to all the citations received by the paper. However, over the years, there emerged several ways to game the system to purposely boost the h-indices/citation counts of a journal or an author (self-citations, citation sharing among co-authors, coercive citations, etc.). For a particular research article, not all citations are equal. Some are significant, while the majority of others are just background or



### 1.3 Chapter Summary

---

contextual. By significant citations, we mean those citing papers which have extensively used the cited paper. Here in this investigation, we are intrigued to see if we could classify citations into significant and contextual citations with the help of ML/NLP and Information Retrieval measures. We analyze the context in which the citation appears in the citing paper, frequency of the in-text citation, citation span, the sentiment of the citances, cross-references, bibliography, similarity in objectives, etc. The study has many implications ranging from accelerating relevant literature discovery, finding insignificant or coercive citations, finding how pervasive a research is in the community via significant citations, establishing a lineage of research, posing a qualitative measure of research quality.

In peer review, this would assist the reviewers to identify the important works necessary to review a paper. Reviewers often struggle with finding the prior influential works that are relevant for a given paper under review (especially if the reviewer is not an expert on the topic). Also, reviewers can identify the papers which might have inspired the current work and also the gradual evolution of the particular research. This would help the reviewers to understand how the work under consideration have helped to take the body of scientific knowledge forward and thereby identify the *novelty* in the current work.

### 1.3 Chapter Summary

In this chapter, we discuss the background and motivation of this thesis. We briefly introduce the problems under investigation and lay the roadmap of the thesis. Finally, we conclude in the final chapter with our future directions of investigation. We enlist our publications at the end of this report.



## CHAPTER 2

---

# Literature Review

---

In this chapter, we survey the prior works that helped to shape our investigation for the various problems we discuss in Chapter 1. We group the literature review according to the problems.

---

## 2.1 Literature Review on Textual Novelty Detection

Textual novelty detection has a history of earlier works with a gradual evolution via different shared tasks. Novelty detection flourished with several efforts in the information retrieval domain. The majority of the works concentrated on extracting sentences that contain novel information with respect to a set of document collections already seen by the system. Our literature survey suggests that although significant efforts are directed towards sentence-level novelty mining, the document-level variant of the problem received comparatively lesser attention. We review the gradual evolution of the problem in the following section.

### 2.1.1 Sentence-level Novelty Detection

We trace the first significant concern on novelty detection back to the new event/first story detection task of the Topic Detection and Tracking (TDT) campaigns [73]. Techniques in TDT mostly involved grouping the news stories into clusters and then measuring the belongingness of an incoming story to any of the clusters based on some preset similarity threshold. If a story does not belong to any of the existing clusters, it is treated as the first story of a new event, and a new cluster is started. Vector space model, language model, lexical chain, etc., were used to represent each incoming news story/document. Some notable contributions in TDT are from Allan et al. (1998) [74], Yang et al. (2002) [75], Stokes and Carthy (2001) [76], Franz et al. (2001) [77], Allan et al. (2000) [78], Yang et al. (1998) [79], Brants et al. (2003) [80]. A close approximation of event-level document clustering via cross-document event tracking can be found in Bagga and Baldwin (1999) [81].

Research on sentence-level novelty detection gained prominence in the novelty tracks of Text Retrieval Conferences (TREC) from 2002 to 2004 [82, 12, 83, 20]. Given a topic and an ordered list of relevant documents, the goal of these tracks was to highlight relevant sentences that contain new information. Significant works on sentence-level novelty detection on TREC data came from Allan et al. (2003) [19], Kwee et al. (2009) [84] and Li and Croft (2005) [17]. Language model measures, vector space models with cosine similarity, and word count measures were the dominant approaches. Some other notable works on finding effective features to represent natural language sentences for novelty computation were based on the sets of terms [85], term translations [86], named entities or NE patterns [87, 88], principal component analysis vectors [89], contexts [18], and graphs [90]. Tsai et al. (2010) [91], Tsai and Chan (2010) [92] presented an evaluation of metrics for sentence-level novelty mining.

## 2.1 Literature Review on Textual Novelty Detection

---

Next came the novelty subtracks in the Recognizing Textual Entailment-Text Analytics Conferences (RTE-TAC) 6 and 7 [93] where textual entailment [94] was viewed as one close neighbor to sentence-level novelty detection. The findings confirmed that summarization systems could exploit the textual entailment techniques for novelty detection when deciding which sentences should be included in the update summaries.

The sentence-level novelty detection methods discussed as above do not scale well to the document-level variant of the problem. The reason being *semantic compositionality*. While the main goal of these methods were retrieval, they do not manifest the composition properties that are essential to comprehend document semantics. Hence, although these approaches helped in our thought process but were not sufficient to address our document-level novelty detection needs. The problem was far from the realization of these IR-based approaches.

### 2.1.2 Document-level Novelty Detection

At the document level, pioneering work was conducted by Yang et al. (2002) [75] via topical classification of online document streams and then detecting novelty of documents in each topic exploiting the named entities. Zhang et al. (2002) [3] viewed novelty as an opposite characteristic to redundancy and proposed a set of five redundancy measures ranging from the set difference, geometric mean, distributional similarity to calculate the novelty of an incoming document with respect to a set of documents in the memory. They also presented the first publicly available Associated Press-Wall Street Journal (APWSJ) news dataset for document-level novelty detection. Tsai and Zhang (2011) [95] applied a document to sentence-level (d2s) framework to calculate the novelty of each sentence in a document which aggregates to detect novelty of the entire document. Karkali et al. (2013) [96] computed novelty score based on the inverse document frequency scoring function. Verheij et al. (2012) [97] presented a comparative study of different novelty detection methods and evaluated them on news articles where language model-based methods performed better than the cosine similarity-based ones. More recently, Dasgupta and Dey (2016) [98] conducted experiments with information entropy measure to calculate the *innovativeness* of a document. Zhao and Lee (2016) [13] proposed an intriguing idea of assessing a user’s novelty appetite based on a curiosity distribution function derived from curiosity arousal theory and Wundt curve in psychology research.

Like the sentence-level novelty detection methods, the document-level novelty detection methods in the literature are not scalable for large document collections. All of them are rule/formula-based IR methods and do not really focus on training a system on *textual novelty* and *non-novelty*. From the inception, we were motivated towards developing supervised approaches that would learn to identify patterns of novelties and redundancies across documents

which were not addressed in the earlier methods. Our approaches aims to bridge the gap between IR-based methods and automatic feature extraction via scalable deep neural architectures for the problem in hand.

### 2.1.3 Diversity Detection

Novelty detection is also studied in information retrieval literature for content diversity detection. The idea is to retrieve relevant yet diverse documents in response to a user query to yield better search results. Carbonell and Goldstein (1998) [99] were the first to explore *diversity* and *relevance* for novelty with their Maximal Marginal Relevance measure. They showed the dichotomy in realizing novelty and relevance simultaneously. Whereas relevance is closer to similarity, novelty is at the opposite end of similarity. However, as we discussed earlier there is no point in realizing novelty for texts which are not relevant. Hence we implicitly manifested the *relevance* criteria for *novelty* when we developed our document-level novelty detection dataset. Some other notable works along this line are from Chandar and Carterette (2013) [100], Clarke et al. (2008) [101], Clarke et al. (2011) [102].

Our investigation significantly differs from the existing literature as we attempt a machine learning classification-based perspective to the problem. Specifically, we investigate models to classify a document into *Novel*, *Non-Novel* classes based on its information content against a set of preceding documents already seen by the system. We further attempt to automatically quantify the *new* information content in documents.

## 2.2 Literature Review on Scope Detection

As discussed, relevance is one important characteristic closely tied with novelty detection. Carbonell and Goldstein (1998) [99] suggested that for deciding on the novelty of the text, we need to consider whether the target text is at all relevant or not. A parallel could be drawn with academic publications. The editor usually judges the scope or relevance of the candidate manuscript with the venue concerned. With our extensive feature engineering (see Chapter 4) and from literature, [67] we see that around 25-30% of the manuscripts are desk-rejected only because they are not relevant or *out-of-scope* of the venue concerned. *Scope Detection* of academic manuscripts is a formal part of the peer-review process.

There have been discussions on implications of AI in peer review by researchers, publication

## 2.2 Literature Review on Scope Detection

---

houses<sup>1</sup> and the meta-science community<sup>2</sup>. However, to the best of our knowledge, we are the first to explore the problem of article classification under the light of scope detection. The idea is to assist both authors and editors in ascertaining the belongingness of the manuscript to the concerned venue based on the recent domain of operation of the journal/conference.

### 2.2.1 Academic Recommender Systems

Our investigation comes close to journal recommendations for academic manuscripts. Most of the journal recommender systems only consider the *Title* and *Abstract* of the paper for generating a suggestion of potential journals where the author may consider to submit her work. Journal publishers have their systems that suggest relevant journals to an author against her work. Examples could be given of Journal Finder by Elsevier<sup>3</sup>, Springer Journal Suggester<sup>4</sup>, EDANZ Journal Selector<sup>5</sup>, etc. Elsevier Journal Finder [103] extracts noun phrases from the paper and matches these with a database using the Okapi BM25 algorithm. However, most publishers and companies provide them to promote their own portfolio or other services. Thus, they diminish the variety of recommendations by only considering their own journals.

Many of the services are black boxes without any information on how they perform their recommendations. There are a few exceptions to this: Journal/Author Name Estimator (JANE)<sup>6</sup> uses the open-source search engine software Lucene to find the 50 most similar papers according to the Lucene index and recommends the journals that occur most often in this set. Also some related web-services eTBLAST [104], GoPubMed [105], HubMed [106], Pubfinder [107], etc. suggest relevant biomedical literatures from PubMed<sup>7</sup> or MEDLINE<sup>8</sup> databases upon user query (typically the title and abstract of the article for which the user wants to find a suitable journal). These systems mostly rely on domain-specific vocabulary matches between the prospective article and different journals to generate a suitable match. Users generally have to submit their article title, abstract, and/or keywords to get a list of potential journals where they could submit their article. There had been quite a lot of work on venue recommendation systems for academic manuscripts. Mention may be made of some notable works by Alhoori and Furuta (2017) [108], Boukhris and Ayachi (2014) [109], Luong et al. (2012) [110], Yu et al. (2018) [111], Chen et al. (2015) [112].

Our problem under consideration is slightly different and mostly targeted towards assisting

---

<sup>1</sup><https://www.wired.com/2017/02/ai-can-solve-peer-review-ai-can-solve-anything/>

<sup>2</sup>[http://events.biomedcentral.com/wp-content/uploads/2017/04/SpotOn\\_Report\\_PeerReview-1.pdf](http://events.biomedcentral.com/wp-content/uploads/2017/04/SpotOn_Report_PeerReview-1.pdf)

<sup>3</sup><http://journalfinder.elsevier.com/>

<sup>4</sup><http://journalsuggester.springer.com/>

<sup>5</sup><https://www.edanzediting.com/journal-selector>

<sup>6</sup><http://jane.biosemantics.org/>

<sup>7</sup><https://www.ncbi.nlm.nih.gov/pubmed/>

<sup>8</sup><https://www.nlm.nih.gov/bsd/pmresources.html>

the editors/chairs and authors to let them identify potential *out-of-scope* submissions. In our experiments, we consider every possible channel of information in a research article (text, image, bibliography) to classify a manuscript as *in-scope* or *out-of-scope*. It is clear that scope detection as a problem has not yet been studied exclusively in the scholarly document processing literature.

While ascertaining the scope of a manuscript, we also consider the multimodal information in papers (texts and images). Multimodal deep learning from texts, images, and videos is a popular NLP problem and is widely explored in the works of Poria et al. (2016) [113], Poria et al. (2015) [114].

Majority of the existing literature which looks into finding the belongingness of an article to the scope of a journal are from the point of view of recommender systems. As mentioned, those systems use the title and abstract of the article to find its scope. Our motivation for this research was always from the point of view of an article classification system where we wanted to know the appropriateness of a certain research article to a given venue. For that, it made sense to go beyond the metadata information of research papers and include paper full-text, images, bibliography in our investigations. Later in this thesis, we will see that bibliography/reference section in papers play a crucial role in ascertaining its domain. Our research differs from existing literature in the way we define the problem as an article classification one; not build another article recommendation system. Also in our solution approaches we encompass full paper discourse in our experiments which were not obvious in earlier systems.

## 2.3 Literature on AI for Peer Review

Artificial Intelligence in academic peer review is a less explored territory. However, with the recent progress in AI research, the topic is gradually gaining attention from the community. There are many use-cases where AI may find application in the peer review process and scholarly communications in general, ranging from reviewer recommendation, venue recommendation, scope detection, novelty detection, plagiarism detection, article quality assessment, citation analysis for article quality, measuring academic influence, etc. Two of the important problems we are investigating as part of this thesis as detailed in Chapter 3 and Chapter 4. This section mostly focuses on the literature for AI in the peer review system and prior works addressing peer-review quality.

A thorough study of the various means of computational support to the peer review system is made by Price and Flach (2017) [115]. Mrowinski et al. (2017) [48] explored an evolutionary algorithm to improve editorial strategies in peer review. The famous Toronto Paper Matching system [49] was developed to match paper with reviewers. Authors in [116] explored a multi-



## 2.3 Literature on AI for Peer Review

---

instance learning framework for sentiment analysis from the peer review texts. The PeerRead dataset [5] is a useful resource to study the peer review texts for both accepted and rejected papers. Authors in PeerRead also presented simple baseline models to predict the peer review outcome.

We differ from the earlier literature as we intend to extract features from paper full-text, reviews, reviewer’s sentiment to predict the recommendation score and final decision on the fate of a paper in the peer review process. The motivation is to have an AI-assisted system for the area-chair or program-chair to help in their decision-making process.

### 2.3.1 Peer Review Quality

Peer Review Quality has been an important research topic in the Meta Science community since the inception of the Peer Review Congress in 1989<sup>9</sup>. Here in this section, we will brief some specific studies dedicated to peer review quality. Justice et al. (1998) [117] studied a randomized control trial to see the effect of masking author identity to improve peer review quality. The author concluded that masking reviewers to author identity as commonly practiced does not improve quality of reviews. Schroter et al. (2004) [118] studied the effects of training on the quality of peer reviews. The authors found that short training packages have only a slight impact on the quality of peer review. The value of longer interventions needs to be assessed. Jefferson et al. (2002) [119] developed approaches to measure the quality of editorial peer reviews. They opined that editorial peer review, although widely used, is largely untested and its effects are uncertain. The Review Quality Instrument (RQI) was proposed by Rooyen et al. (1999) [120] to assess peer reviews of manuscripts. The 8-point Review Quality Instrument (RQI) consisted of qualitative questions on the overall review and had questions like *Were the reviewers’ comment constructive?*, *Did the reviewer discussed the originality of the paper?*, etc. The total score is calculated as the mean of the first 7 items, while the 8th “global item” provides an extra validation check. Shattell et al. (2010) [121] studied the author’s and editor’s perspective on peer review quality in three scholarly nursing journals. They concluded that it is incumbent upon editors and reviewers to provide guidance and support to reviewers. Manuscript reviews could be improved by increasing the consistency of numeric ratings, narrative comments, and recommendations regarding disposition of the manuscripts. Rooyen et al. (2001) [122] proposed an evaluation framework for peer review quality. They detail the research into the quality of peer review, in particular the BMJ’s programme of research to date, the results obtained, and consequent changes in practice. A randomized control trial to see how mentoring new peer reviewers to improve review quality was done by Houry et al. (2012) [123]. The

---

<sup>9</sup><https://peerreviewcongress.org>

authors found the structured training intervention of pairing newly recruited medical journal peer reviewers with senior reviewer mentors did not improve the quality of their subsequent reviews. A systematic review and meta-analysis on the impact of interventions to improve the quality of peer reviews of biomedical journals were conducted by Bruce et al. (2016) [124]. Enserink (2001) [125] explored the dubious connection between peer review and quality. In their meta-analyses they found that there is little evidence that peer review actually improves the quality of research papers. D’Andrea and O’Dwyer (2017) [126] argued if editors can save peer reviews from peer reviewers They found that that the biggest hazard to the quality of published literature is not selfish rejection of high-quality manuscripts but indifferent acceptance of low-quality ones. Thirty years on from the first congress on peer review, Drummond Rennie [127] reflects on the improvements brought about by research into the process — and proposed directions to make the peer review process scientific. Callaham et al. (1998) [128] investigated the reliability of the editor’s subjective quality ratings of peer review of manuscripts. They concluded that subjective editor ratings of individual reviewers were moderately reliable and correlated with reviewer ability to report manuscript flaws. Sizo et al. (2019) [129] provides an overview of assessing the quality of peer review reports of scientific articles. Sculley et al. (2019) [130] proposed a rubric to hold reviewers to an objective standard for review quality. Superchi et al. (2019) [131] presents a comprehensive survey of criteria tools used to assess the quality of peer review reports in the biomedical domain. Wicherts (2016) [132] proposed that the peer-review process’s transparency may be seen as an indicator of the quality of peer-review and developed and validated a tool enabling different stakeholders to assess the transparency of the peer-review process. Quality in peer review is an active area of research within the peer review, meta-research, and scholarly communication communities (especially in the biomedical domain) with focused events like Peer Review Week<sup>10</sup>, Peer Review Congress, and COST Action PEERE New Frontiers of Peer Review consortium<sup>11</sup>.

Rooyen et al. (1999) [133] studies the effect of revealing the identity of the reviewer in the peer review process and if it can enhance the quality of peer reviews. In contrast, Justice et al. (1998) [117] studies the effect of the blind review model. Xiong and Litman (2011) [134] predicts the helpfulness of a peer-review using manual features.

However, our investigation differs from these works in the sense that we attempt to see whether a deep model could predict the peer review outcome or comment on the quality of the peer reviews. None of the existing literature on peer review quality attempts to derive a computational method to quantify the quality of peer reviews. One reason being that peer

---

<sup>10</sup><https://peerreviewweek.wordpress.com>

<sup>11</sup><https://www.peere.org/>

## 2.3 Literature on AI for Peer Review

---

review data are highly confidential due to proprietary reasons and also annotated data for review quality is not available in the public domain. Our work tries to bridge this gap and is a first of its kind to apply NLP/ML methods to *rate the reviewers*.

### 2.3.2 Literature Review on Finding a Research Lineage

We pursue a novel problem to establish a research lineage of a given piece of research. Finding the research lineage would also help identify how influential a given paper is in the community and how many different works it has inspired. Thus the problem is closely related to finding the influence of a work in academia. We try to attempt this via identifying *significant* citations.

Measuring academic influence has been a research topic since publications associate with academic prestige and incentives. Several metrics (Impact Factor, Eigen Factor, *h*-index, citation counts, altmetrics, etc.) came up to comprehend research impact efficiently. Still, each one is motivated on a different aspect and has found varied importance across disciplines. Zhu et al. (2015) [135] did pioneering work on academic influence prediction leveraging on citation context. Shi et al. (2019) [136] presented a visual analysis of citation context-based article influence ranking. Xie et al. (2016) [137] predicted paper influence in an academic network by taking into account the contents and venue of a paper, as well as the reputation of its authors. Shen et al. (2016) [138] used topic modeling to measure academic influence in scientific literature. Manju et al. (2017) [139] identified influential researchers in an academic network using a rough-set based selection of time-weighted academic and social network features. Pileggi (2018) [140] did a citation network analysis to measure academic influence. Zhang and Wu (2020) [141] used a dynamic academic network to predict the future influence of papers. Tang and Chen (2019) [142] analyzed the impact of academic papers based on improved PageRank. Wang et al. (2019) [143] assessed the academic influence of scientific literature via altmetrics. Zhao et al. (2019) [144] measured academic influence using heterogeneous author-citation networks.

The closest literature for our problem are the ones on citation classification. Citation classification has been explored in the works of Dong and Schafer(2011) [145], Teufel et al. (2006) [146], Alvarez et al. (2017) [147], Qayyum and Afzal (2019) [148]. More recently, several open-source datasets for citation classification came up in the works of Cohan et al. (2019) [149], Pride and Knoth (2020) [150]. Valenzuela et al. (2015) [151] explored citation classification into *influential* and *incidental* using machine learning techniques. However, our problem is motivated beyond citation classification, and to the best of our knowledge, we did not find any work leveraging citation classification for finding a research lineage.

In peer review, this would help the reviewers identify the important works necessary to review a paper. Reviewers often struggle with finding the prior influential works that are relevant for

a given paper under review (especially if the reviewer is not an expert on the topic). Also, reviewers can identify the papers which might have inspired the current work and even the gradual evolution of the particular research. This would help the reviewers to understand how the work under consideration has helped to take the body of scientific knowledge forward and thereby identify the *novelty* in the current work.

Measuring academic influence is an important research topic in the scientometrics and *science of science* communities where researchers have used paper meta data to predict the influence of a research paper. We were motivated to go beyond paper meta data and apply NLP techniques to mine citation contexts from paper full-text to predict the importance of a given citation. Also, as we mention that our problem is motivated beyond simple citation classification, we investigated to see how influence propagation in a citation network would help to automatically generate a *research lineage* for a given topic.

## 2.4 Chapter Summary

In this chapter, we discuss the relevant works that inspired our investigation for textual novelty detection. We reviewed relevant works for sentence-level and the document-level variant of the problem. We limit our exploration of novelty to the newspaper domain and hope to explore the problem in the scholarly domain in the future.

We investigated the rest of our problems on scholarly text mining. We reviewed the literature on venue recommendation for scope detection as it is the closest task to the problem. Next, we presented an extensive review of relevant literature for AI in Peer Review and Peer Review Quality. Finally, we discuss our literature study on academic influence and citation classification for our investigation to establish a research lineage for a given work.

## CHAPTER 3

---

# Textual Novelty Detection

*Of all the passions of mankind, the love of novelty most rules the mind.*

*—Shelby Foote*

---

In this first contribution chapter, we detail our investigations on textual novelty detection. Due to the lack of gold-standard data on document-level novelty, we build a dataset from newspaper articles. We discuss several approaches towards classifying a document based on its new information content and quantifying textual novelty. Finally, we show that multi-premise entailment is one very close approximation towards semantic-level non-novelty. We look forward to continuing the document-level investigations enlisted here to the scholarly article texts in the future.

---

### 3.1 Introduction

The quest for new information is an inborn human trait and has always been the quintessential skill for human survival and progress. Novelty drives curiosity, which in turn drives innovation. Most of the breakthrough discoveries and remarkable inventions throughout history, from flints for starting a fire to self-driving cars, have something in common: they are the result of curiosity. The impulse to seek new information and experiences and explore novel possibilities is a basic human attribute. Humans elicit novel signals from various channels: text, sound, scene, via basic senses, etc. Novelty is important in our lives to drive progress, to quench our curiosity needs. Arguably the biggest source of information elicitation in this age of digitization are texts: be it the book, the web, the papers, etc. However with the abundance of information, comes the problem of duplicates, near-duplicates, and redundancies. Although, document duplication is encouraged in certain use-cases, it impedes the search for new information. Hence identifying redundancies is important to seek novelties. We humans are already equipped with an implicit mechanism (*Two Stage Theory of Human Recall*: recall-recognition [152]) via which we can segregate new information from old ones. In our work we are interested to explore how machines would be able to identify semantic-level non-novel information and hence pave the way to identify documents having significant content of new information. Specifically, here in this work, we investigate how can we automatically discover novel knowledge from the dimension of text or rather how can we identify that a given text has new information. We rely on certain principles of Machine Learning and Natural Language Processing to design efficient neural architectures for textual novelty detection at the document-level.

#### 3.1.1 Why Textual Novelty Detection is important?

In Natural Language Processing (NLP), Novelty Detection refers to the task of finding texts that has some new information to offer with respect to whatever is earlier seen or known. With the exponential growth of information all across the web, there is an accompanying menace of redundancy. A considerable portion of the web contents are duplicates, and we need efficient mechanisms to retain new information and filter out redundant ones. However, detecting redundancy at the semantic level and identifying novel texts is not straightforward because the texts may have less lexical overlap yet convey the same information. On top of that, non-novel/redundant information in a document may have assimilated from multiple source documents, not just one. The problem surmounts when the subject of the discourse is documents, and multiple prior documents need to be processed to ascertain the novelty/non-novelty of the current one in concern. With the exponential rise of information across the web, the problem becomes more relevant

### 3.1 Introduction

---

now as information duplication (prevalence of non-novel information) is more prominent. The deluge of redundant information impedes critical, time-sensitive, and quality information to the end-users. Duplicates or superfluous texts hinders the reach to new information that may prove crucial to a given search. According to a particular SEO study<sup>1</sup> by Google in 2016, 25-30% of documents on the web exists as duplicates (which is quite a number!). With the emergence of humongous language models like GPT3 [153], machines are now capable of generating artificial and semantically redundant information. Information duplication is not just restricted to lexical surface forms (mere copy), but there is duplication at the level of semantics [154]. Hence, identifying whether a document consists of new information to the reader’s interest is a significant problem to explore so as to save space, time, and retain attention of the reader. Novelty Detection in NLP finds application in several tasks, including text summarization [155], plagiarism detection [156], modeling interestingness [157], tracking the development of news over time [2], identifying fake and misinformation [158], etc.

#### 3.1.2 Our Exploration on Textual Novelty Detection

Textual novelty detection is known for long as an information retrieval problem [159] where the goal is to retrieve relevant pieces of text that carry new information with respect to whatever is previously seen or known to the reader. Novelty detection as an information retrieval problem signifies retrieving relevant sentences that contain new information in a discourse. Sentence-level novelty detection[159] although important, but would not suffice in the present-day deluge of web information in form of documents. Hence, we emphasize on the problem’s document-level variant which is to categorize a document (as novel, non-novel, or partially novel) based on the amount of new information in the concerned document. Sentence-level novelty detection is a well-investigated problem in information retrieval [17, 101, 12, 82]; however, we found that document-novelty detection attracted relatively less attention in the literature. Moreover, the research on the concerned problem encompassing semantic-level comprehension of documents is scarce. Maybe because of the argument that every document contains something in new [20]. Comprehending the novelty of an entire document with confidence is even a complex task for humans. Robust semantic representation of documents is still an active area of research, which also somewhat limits the investigation of novelty mining at the document-level. Hence, categorizing a document as novel or non-novel is not straightforward and involves complex semantic phenomena of inference, relevance, diversity, relativity, temporality, as we show in our work[2].

---

<sup>1</sup><https://searchengineland.com/googles-matt-cutts-25-30-of-the-webs-content-is-duplicate-content-thats-okay-180063>

Here in this dissertation work, we cast the document-level variant of the problem as a classification problem in machine learning. We aim to classify a document as *novel* or *non-novel* based on the *new* information content in the document. Although we started our explorations on this problem with scholarly data (research articles), we quickly realized that *defining and finding scientific novelty from texts is not straightforward*. Also, if we are to attempt the problem we would need high-quality gold-standard data from domain-experts. Deciding whether a scientific text is universally *novel* or not is relative, time-dependent, and subjective. There are no datasets that would cater to this domain of the problem and we presume that it would be an effort of epic proportions to curate a real-world scientific novelty dataset. Scientific texts are *intelligent* texts and have implicit background information, the source of those could not be traced easily. Even if traced, reasoning over the background information to arrive at a conclusion in sync with the current text in concern is not straightforward.

Hence, we started working with objective texts: newspaper articles. However, except the AP-WSJ dataset [160], there were no document-level novelty detection datasets available. Even the APWSJ dataset was developed from an information retrieval perspective and only 8.9% of its documents are *non-novel*. Hence we decided to curate our own dataset from newspaper articles to facilitate our machine learning experiments. We then went on to develop several methods to address textual novelty detection from a classification perspective.

## 3.2 TAP-DLND 1.0 corpus

We name our dataset after the three primary investigators: *Tirthankar-Asif-Pushpak Document-Level Novelty Detection* (TAP-DLND) [8]. The dataset is balanced and consists of 2736 *novel* documents and 2704 *non-novel* documents. There are several categories of events; ten to be precise (Business, Politics, Sports, Arts and Entertainment, Accidents, Society, Crime, Nature, Terror, Society). For each novel/non-novel document, there are three source documents against which the target documents are annotated. While developing this dataset we ensured that *Relevance, Relativity, Diversity, Temporality* [8] characteristics are preserved.

### Relevance

The target document should be relevant to prior knowledge. For example, seeking novelty between two documents, one talking about *jaguar*, the animal and the other about *jaguar*, the car is futile as one is not relevant to the other. Quite obvious that each one would contain different information than the other. So *Relevance* should hold.



#### Diversity

*Diversity* correlates with the new information content. More the new information in a document, diverse would be the content. Hence novel information should be relevant yet *diverse* from existing information.

For example, let us consider, on a given date a certain newswire document  $X$  reports about an accident at a certain place. On the subsequent date another reporting  $X'$  surfaces which details about the investigation being carried out by police. Now  $X'$  will contain new information with respect to  $X$ . That is to say, given a reader has already read about the first reporting (facts of the accident)  $X$ , the second reporting  $X'$  having significant different content as well as different direction of reporting (or intent) would appear novel to the reader. So  $X'$  is relevant to  $X$  yet divergent i.e. containing new information.

#### Relativity

The amount of new information content is important while deciding the novelty of an entire document. When we talk about a document being novel it is always with respect to a reference set of documents already seen (information already gained from those seen documents) or what we say as the knowledge base of the reader. So the quantity of *relative* new information plays a role for deciding document novelty.

#### Temporality

Finally novel information is usually a temporal update over existing knowledge. The previous example justifies the view.

With these notion of novelty we went on to create a resource that effectively taps these properties, *viz.*, **Relevance**, **Diversity**, **Relativity** and **Temporality**. Our resource not only encompasses the lexical form of redundancy (a straight forward form of *non-novelty*) but also delves deep into semantic textual redundancy (a more complex form of *non-novelty*). For developing this resource, we tracked the development of an event (news items) across time over several Indian newspapers. We did a temporal crawling of event-specific news items published by different newspapers over a specific period. For a particular event, we select a set of documents as the *source* knowledge or the *prior relevant knowledge* of the event and the rest as *target* documents (for which the state of novelty would be ascertained). The core idea is: for a given event (*e.g.*, *reporting of an accident in Bali*), different newspaper would report more or less similar content on a given date. On the subsequent dates, new information regarding the event may surface up (*e.g.*, *the accident*

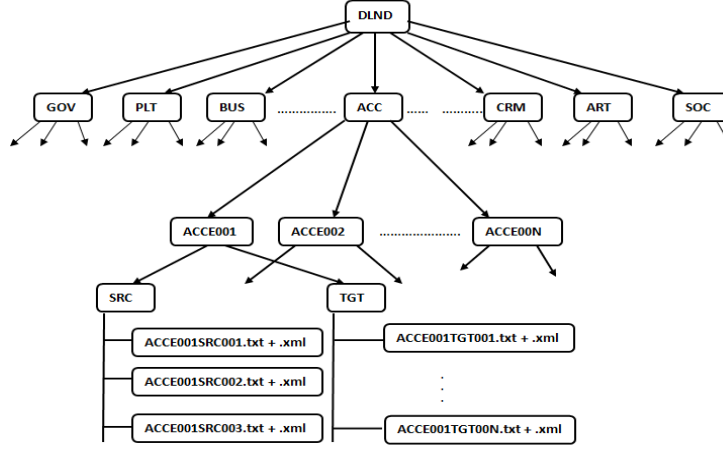


Figure 3.1: The TAP-DLND 1.0 corpus structure. We retain the structure in the extended dataset we use in the current work

was a *actually plot*). The relevant temporal information update over the existing knowledge is what we deem as *novel knowledge*. We intentionally choose such events which continued in the news for some days to facilitate our notion of novelty update. We ask our annotators to judge the information content in the target document against the source documents only [Annotation Label: NOVEL or NON-NOVEL]. Since we chose the source and target documents from the same event, hence we preserve the *relevance* property. We annotate the target document based on its diversity of information, hence we encompass *diversity*. The more a target document has diverse information, the more is its affinity towards novelty. The *relative* amount of new information in the target document is the subjective criteria that makes a document appear *novel* or *non-novel* to the annotator. We leave it to the annotator judgement to decide on the label of the target document with respect to the source documents. However, via majority voting we arrive to a consensus to decide on the label of the particular candidate document. We select the target documents which are published at a later date than the source document in order to hold the *temporality* criteria. We follow the following *annotation principles*:

1. To annotate a document as *non-novel* whose semantic content significantly overlaps with the source document(s) (maximum redundant information). This was subjective to the annotator to determine if the amount of *newness* in the target document is sufficient to label it as *novel*. We instructed that to label a document as *novel*, the target document should contain updates over the existing source information on the event and the central theme of the target document should be on the information update. *Non-novel* documents should report almost the same information that are present in the source documents. Almost each target document may contain some additional information but as long as the central information remains the same as the source ones, we label those target documents

### 3.3 TAP-DLND 2.0 corpus

---

as *non-novel*.

2. To annotate a document as *novel* if its semantic content as well as intent (direction of reporting) significantly differs from the source document(s) (minimum or no information overlap). It could be an update on the same event or describing a post-event situation.
3. We left out the ambiguous cases (for which the human annotators were not sure about the label).

Due to the subjective nature of the annotations, we performed the annotations twice, and resolved the disagreements with the third annotator (the author). Our dataset manifests the presence of semantic-level redundancies, goes beyond lexical similarity, and hence it makes an ideal candidate for our experiments. With respect to the chosen source documents, we found novel documents appearing in later dates of the event in chronological order and the non-novel documents are found from the initial days of the event reporting (usually the dates from which the source documents are selected). The inter-rater agreement is 0.82 in terms of Fleiss Kappa [1], and the average length of documents is 15 sentences/353 words. Figure 3.1 shows the organization of our dataset.

### 3.3 TAP-DLND 2.0 corpus

We extend our TAP-DLND 1.0 corpus in the 2.0 version. Whereas TAP-DLND 1.0 is for document-level novelty classification, the TAP-DLND 2.0 dataset is catered towards deducing the novelty-score of a document (quantifying novelty) based on the information contained in the preceding/source documents. We annotate the new dataset at the sentence-level (more fine-grained) in an attempt to weed out inconsistencies that may have persisted with document-level annotations. We re-annotate TAP-DLND 1.0 from scratch, now at the sentence level, extend

Table 3.1: TAP-DLND 2.0 and TAP 1.0 dataset statistics. Inter-rater agreement [1] is measured for 100 documents for sentence-level annotations by two raters.

Dataset Characteristics	TAP 2.0	TAP 1.0
Event categories	10	10
Number of events	245	223
Number of source documents per event	3	3
Total target documents	7536	6109
<b>Total annotations</b>	<b>120,116 sentences</b>	5400 documents
Average number of sentences per document	~ 16	~15
Average number of words per document	~ 385	~353
Inter-rater agreement	0.88	0.82

to more than 7500 documents and finally deduce a document-level novelty score for each target document. The judgment of novelty at the document-level is not always *unanimous* and

is subjective. Novelty comprehension also depends on the appetite of the observer/reader (in our case the annotator or the labeller) [13]. Also, it is quite likely that every document may contain something in new with respect to previously seen information [12]. However, this relative

The interface is titled "My Splitter" and includes three buttons: "Upload the .txt file", "Open the sources", and "Edit Target file". Below these is a table with two columns: "Sentence" and "Feedback".

Sentence	Feedback
Twelve people were killed as seven coaches and the engine of the Jagdalpur-Bhubaneswar Express derailed near Kureru station in Vizianagaram district of Andhra Pradesh during the intervening night between Saturday and Sunday .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
The incident took place around 11 pm when the train was going to Bhubaneswar .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
" Seven coaches and the engine of the 18448 Jagdalpur-Bhubaneswar Express derailed near Kureru station .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
Besides the engine , the luggage van , two general coaches , two sleeper coaches , one AC three tier coach and an AC two tier coach derailed , "	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
Chief PRO of East Cost Railway J P Mishra said .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
" According to doctors at the site , twelve people have died in the incident , " he said .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
" Four accident relief vans have been sent to the accident site .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
The reason behind the incident is yet to be ascertained , " the official said .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED

Figure 3.2: The Sentence-Level Annotation Interface used to generate the Document-Level Novelty Score (Gold Standard)

amount of new information is not always justified to label the entire document as novel. Also, significance of the *new* information with respect to the context plays a part. It may so happen that a single information update is so crucial and central to the context that it may affect the novelty comprehension of the entire document for a labeller. Hence, to reduce inconsistencies, we take an objective view and deem that instead of looking at the target document in entirety, if we look into the sentential information content, we may get more fine-grained new information content in the target document discourse. Thus with this motivation, we formulate a new set of annotation guidelines for annotations at the sentence-level. We associate scores with each annotation judgment which finally cumulates to a document-level novelty score. We design an easy to navigate interface (Figure 3.2) to facilitate the annotations and perform the annotation eventwise. For a particular event, an annotator reads the predetermined three seed source documents, gathers information regarding that particular event and then proceeds to annotate the target documents, one at a time. Upon selection of the desired target document, the interface splits the document into constituent sentences and allows six different annotation options for each target sentence (Table 3.2). We finally take the cumulative average as the document-level novelty score for the target document. We exclude the sentences marked as irrelevant (IRR) from the calculation. The current data statistics for TAP-DLND 2.0 is in Table 3.1.

### 3.3 TAP-DLND 2.0 corpus

---

Table 3.2: Sentence-level annotations. The target document sentences are annotated w.r.t. the information contained in the source documents for each event. The annotations are qualitatively defined. We assign scores to quantify them.

Annotation Labels	Description	Score
Novel (NOV)	The entire sentence has new information.	1.00
Non-Novel (NN)	The information contained in the sentence is completely redundant.	0.00
Mostly Non-Novel (PN25)	Most of the information is overlapping with the source with little new information.	0.25
Partially Novel(PN50)	The sentence has an almost equivalent amount of new and redundant information.	0.50
Mostly Novel (PN75)	Most of the information in the sentence is new.	0.75
Irrelevant (IRR)	The sentence is irrelevant to the event/topic in context.	—

#### Annotator Background

Two annotators with post-graduate level knowledge in English were involved in labeling the TAP-DLND 1.0 and 2.0 target documents. Since the documents are English newspaper articles, it did not require any special expertise apart from command over English to annotate the target documents. Our annotators were exclusively hired as full-time research staff and were paid equivalent to a Ph.D. student in India. Two annotators independently labeled the target documents. The author resolved the differences via majority voting. We found that novel items with respect to the source documents were mostly found in the reporting published in subsequent dates. Whereas non-novel items we found in the reporting published by different agencies in the same date as that of the source documents. This is in line with the *Temporality* criteria we discussed earlier. The inter-annotator agreement ratio was found to be 0.82 in terms of Kappa coefficient [1] for the document-level annotations (TAP-DLND 1.0) which is assumed to be good. The inter-rater agreement for sentence-level annotations for TAP-DLND 2.0 is 0.88. The duration of annotation was close to 10 months.

### 3.4 Feature-Engineering Approach for Novelty Detection

As our very first exploration, we experiment with some simple similarity and diversity-based measures with the intuition that *novelty is an opposite characteristic of text similarity and closer to diversity* [2]. Table 3.3 displays the definition of our feature-set.

Table 3.3: Feature Set Definition for Novelty Classification [2]

Type	Features	Description
Semantic	Paragraph Vector (pv) + Cosine	We represent the source and target documents in terms of <i>paragraph vectors</i> <sup>2</sup> [161]. Then we take the maximum of the cosine similarity between the source-target pairs.
Semantic	Concept Centrality	To identify the central theme of a document we use the <i>TextRank</i> summarization algorithm by [162]. Thereafter we vectorize the ranked summary for each source and target document by simple <i>word2vec</i> <sup>3</sup> [163] concatenation. Finally we take the maximum of the cosine similarity between the source and target vectors.
Lexical	n-gram similarity	We compute lexical overlap of target <i>n-gram</i> 's with respect to source documents for $n = 2, 3$ and 8. Octagrams we use to put emphasis on phrase overlap.
Lexical	Named Entities and Keywords match (kw-ner)	As Named Entities <sup>4</sup> and Keywords <sup>5</sup> play a significant role in determining <i>relevance</i> , we put additional weightage to them by considering their match (target w.r.t. sources) as a separate feature.
Lexico-Semantic	New Word Count (nwc)	The number of new words could be an effective indicator of the amount of novel information content in the target document w.r.t. the source(s) given. Here, for calculating new words, along with the surface forms, we consider their synonyms <sup>6</sup> as well to establish semantic relatedness.
Language Model	Divergence (kld)	We use this feature to measure the dissimilarity between two documents represented as language models. We concatenate all the source documents into one and then measure the Kullback-Leibler Divergence with the target.

#### Analysis

We investigated the significance of each feature by measuring the Information Gain (See Figure 3.3). The information gain for a feature  $x_k$  is the expected reduction in entropy-i.e., uncertainty achieved by learning the state of that feature. We attribute the better performance of our approach to the choice of semantic features for our experiments (see Figure 3.3). Lexico-Semantic feature new word count has the maximum contribution, for which we argue that novel events in context to newspaper articles would contain new entities, concepts, numbers whereas non-novel documents would consist identical or synonymous entities. Semantic features play a vital role which indicates that detection of novelty extends beyond lexical characteristics of text. Paragraph vector similarity contributes to detect the semantic closeness (manifesting non-novelty) and distance (novelty). Keywords, Named-Entities are important attributes to see if the source-target texts are *relevant* or not. *N-gram* overlap signifies the overlap in terms of tokens, strings. It is interesting to see that semantic similarity has more contribution in the classification

### 3.4 Feature-Engineering Approach for Novelty Detection

---

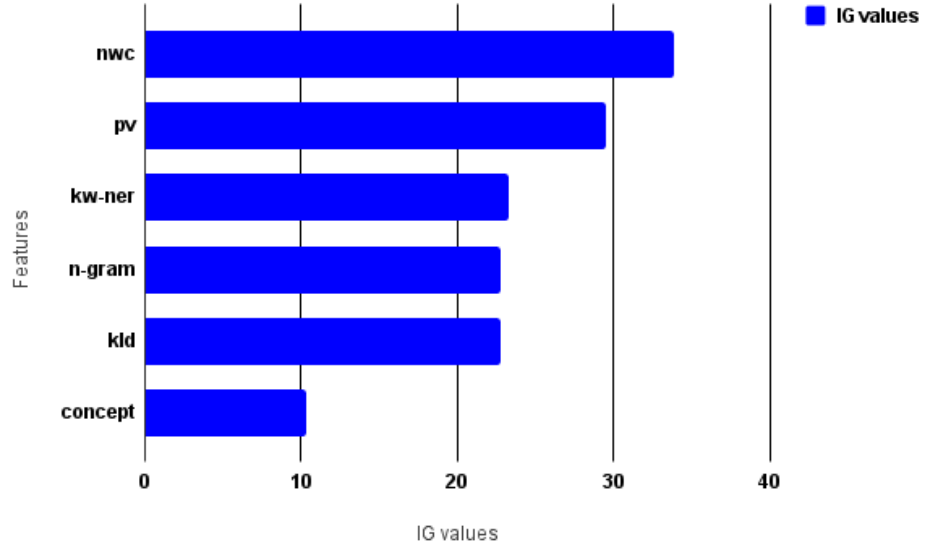


Figure 3.3: Significance of features *based on Information Gain (IG)*. The length of the bar corresponds to the average merit (X : IG) of the feature (: Y).

that lexical overlap signifying that the dataset manifests the semantic-level textual novelty/non-novelty. KL-Divergence measures the diversity of distribution of the language model, performs comparatively inferior than other features. The reason could be the underlying language model which was simple bag of words. Finally the concept centrality feature depends on summarization similarity which probably do not captures all semantic information units within the source-target pairs.

### 3.5 Relative Document Vector-Convolutional Neural Network

In our subsequent exploration, having an acceptable benchmark dataset available at our end, instead of handcrafting the similarity/divergence based features, we try to learn feature representations from a target document with respect to the source document(s) using a Convolutional Neural Network (CNN). The objective is to see how a deep neural network can automatically extract features from documents for novelty classification. Our proposed model is based on a sentence embedding paradigm proposed by [6]. We leverage their idea and create a representation of the *relevant* target document *relative* to the designated source document(s) and call it as the *Relative Document Vector (RDV)*<sup>7</sup>. We then train a Convolutional Neural Network (CNN) with the RDV of the target documents in the similar line of [165], and finally classify a document as *novel* or *non-novel* with respect to its source documents (Figure 3.4). The role of CNN is that of a deep feature extractor to extract meaningful features from the RDV. CNN is widely used in several NLP tasks for faster feature extraction compared to RNNs or LSTMs. Although there are document embedding models available, our method is specifically tailored to address the *relativity* and *diversity* criteria which is fundamental to the definition of *novelty*. Here  $T_1$  is the *target* document whose state of *novelty* is to be determined against the source document(s)  $S_1, S_2, \dots, S_M$  i.e. to say the objective is to automatically figure out whether  $T_1$  is *novel* or not once the machine has already seen/scanned  $S_1, S_2, \dots, S_M$ . Our model assumes that the documents are relevant to a context.

#### 3.5.1 Proposed Method

##### Embedding and Sentence Encoder

The task of *Novelty Detection* requires high-level understanding and reasoning about semantic relationships within texts. Textual Entailment or Natural Language Inference is one such task which exhibits such complex semantic interactions. Following from [6] we therefore employ a sentence encoder based on a bi-directional Long Short Term Memory (LSTM) architecture with max pooling, trained on the large-scale Stanford Natural Language Inference (SNLI) dataset [166]. SNLI entries supposedly captures rich semantic associations between the text pairs (entailment/inference relationships between premise and hypothesis). The output of our sentence encoder is a fixed sized (2048 dimension) sentence embedding of each of the sentences of the input document (source or target).

---

<sup>7</sup>We call our document vector *relative*, as we desire to encode the relative *new* information of a target document w.r.t. its relevant source document(s)



### 3.5 Relative Document Vector-Convolutional Neural Network

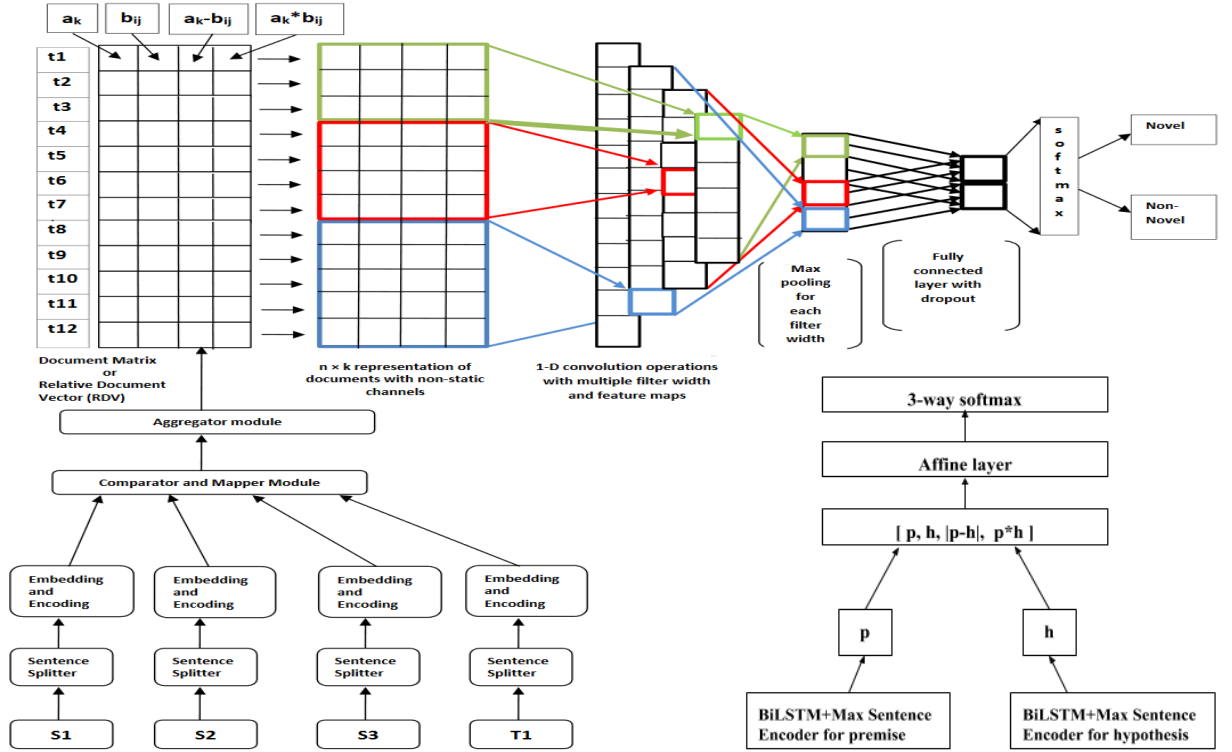


Figure 3.4: RDV-CNN framework for Novelty Detection. Generic SNLI Training [6]. The sentence encoder is trained on SNLI. The RDV-CNN is trained with the respective novelty datasets.

#### BiLSTM with max pooling

For a sequence of  $T$  words  $\{w_t\}_{t=1, \dots, T}$ , a bidirectional LSTM computes a set of  $T$  vectors  $\{h_t\}_t$ . For  $t \in [1, \dots, T]$ ,  $h_t$  is the concatenation of a forward LSTM and a backward LSTM that read the sentences in two opposite directions. We combine the varying number of  $\{h_t\}_t$  to form a fixed-size vector by selecting the maximum value over each dimension of the hidden units (max pooling).

$$\begin{cases} \vec{h}_t = \overrightarrow{LSTM}_t(w_1, \dots, w_T) \end{cases} \quad (3.1)$$

$$\begin{cases} \overleftarrow{h}_t = \overleftarrow{LSTM}_t(w_1, \dots, w_T) \end{cases} \quad (3.2)$$

$$\begin{cases} h_t = [\vec{h}_t, \overleftarrow{h}_t] \end{cases} \quad (3.3)$$

#### Training the sentence encoder

We follow the generic SNLI training scheme of [6]. The semantic nature of SNLI corpus makes it a good candidate for learning universal sentence embeddings in a supervised way to capture universally useful features. The training<sup>8</sup> on SNLI is done using the shared sentence encoder that

<sup>8</sup>SGD with a learning rate of 0.1 and a weight decay of 0.99 are used. At each epoch, the learning rate is divided by 5 if the development accuracy decreases. Training is done with a mini-batch size of 64. GloVe vectors

outputs a representation for the premise  $p$  and hypothesis  $h$  (Figure 3.4). The representations are further concatenated as:

$$[p, h, |p - h|, p * h]$$

to form a resulting vector, which captures information from both the premise and hypothesis, and is fed into a 3-class classifier<sup>9</sup> (multi-layer perceptron with 1 hidden layer of 512 units) consisting of multiple fully-connected layers culminating into a softmax layer.

### Comparator Module

This module finds the closest sentence in the source document(s) with respect to each of the target sentences. Thus each sentence ( $t_k$ ) in a target document ( $T$ ) is mapped to its nearest source sentence ( $s_{ij}$ ) where  $i$  denotes the source document and  $j$  signifies the sentence position within that document. The encoder module outputs a fixed-length vector representation  $a_k$  for the target sentence  $t_k$  and  $b_{ij}$  for any source sentence  $s_{ij}$ . We define closeness as the maximum of cosine similarity between the vectors  $a_k$  and  $b_{ij}$ .

$$s_{cosine}(\vec{a}_k, \vec{b}_{ij}) = \frac{\vec{a}_k \cdot \vec{b}_{ij}}{|\vec{a}_k| |\vec{b}_{ij}|} \quad (3.4)$$

Thus we have a mapping relationship for each sentence in the target document with one of the source sentences.

$$b_{ij} \rightarrow a_k$$

where  $a_k$  has the max. cosine similarity with  $b_{ij}$ .

### Aggregator module

This module aggregates the mappings produced in the comparator module to generate a document matrix. The mapping of a target sentence  $t_k$  to its closest source sentence  $s_{ij}$  is rendered by constructing a feature vector that captures the relation between the source and the target. This feature vector consists of the concatenation of the two sentence embeddings corresponding to  $t_k$  and  $s_{ij}$ , their absolute element-wise difference, and their element-wise product [167]. The first heuristic follows the most standard procedure of the ‘‘Siamese’’ architecture, while the latter two are certain measures of ‘‘similarity’’ or ‘‘closeness’’. Thus, the *Relative Sentence Vector* (RSV) corresponding to a target sentence  $t_k$  is represented as :

$$RSV_k = [a_k, b_{ij}, |a_k - b_{ij}|, a_k * b_{ij}] \quad (3.5)$$

---

trained on Common Crawl 840B (<https://nlp.stanford.edu/projects/glove/>) with 300 dimensions are used as fixed word embeddings.

<sup>9</sup>SNLI has 3 classes of sentence-pair judgments: neutral, contradiction, entailment

### 3.5 Relative Document Vector-Convolutional Neural Network

where comma (,) refers to the column vector concatenation. This representation is inspired from the word embedding studies [168] where the linear offset of vectors is seen to capture semantic relationships between the two words. Authors in [167] successfully leveraged this idea for modeling sentence-pair relationships which we alleviate to model documents. Thus for each target sentence  $t_k$  we compute the RSV and aggregate them to form the *Relative Document Vector* (RDV) of target document  $T_j$  with respect to the source documents(s)  $S_i$ . Aggregation is realized as a *slot filling task* to shape the document matrix<sup>10</sup> or RDV of dimension  $N \times 4D$  where  $N$  is the number of sentences in a target document (padded when necessary) and  $D$  is the sentence embedding dimension produced by the encoder module. *Our rationale behind the RDV-CNN is: The operators: **absolute element-wise difference** and **product** would result in such a vector composition for **non-novel** sentences which would manifest 'closeness' whereas for **novel** sentences would manifest 'diversity'; the aggregation of which would aid in the interpretation of document level **novelty** or **redundancy** by a deep neural network. We chose CNN due to its inherent ability to automatically extract features from distinct representations.*

#### CNN module

The document matrix or the RDV now becomes the input to a CNN<sup>11</sup> for training and subsequent classification of a target document as *novel* or *non-novel* with respect to its source document(s). The CNN component is similar to the one as used in [165] for sentence classification. The notable difference is in the usage of word embedding. Instead of *word2vec* embeddings we use here the relative sentence embeddings of dimension  $4D$  ( $k^{th}$  sentence in the document is represented by an embedding vector  $RSV_k \in \mathbb{R}^D$ ). We experiment with the NON-STATICTEXT channel variant (embeddings gets updated during training) of the CNN.

For each possible input channel, a given document is transformed into a tensor of fixed length  $N$  (padded with *zero-tensors* wherever necessary to tackle variable sentence lengths) by concatenating the relative sentence embeddings.

$$RSV_{1:N} = RSV_1 \oplus RSV_2 \oplus RSV_3 \oplus \dots \oplus RSV_N \quad (3.6)$$

where  $\oplus$  is the concatenation operator. To extract *local features*<sup>12</sup>, convolution operation is applied. Convolution operation involves a *filter*,  $W \in \mathbb{R}^{HD}$ , which is convolved with a window

<sup>10</sup>each row in the document matrix is one *relative sentence vector* corresponding to each target sentence  $t_k$

<sup>11</sup>*tanh* as the activation function, filter windows ( $h$ ) of 3,4,5 with 100 feature maps each, dropout rate ( $p$ ) of 0.5 on the penultimate layer with a constraint on  $l_2$ -norms of the weight vectors. *ADADELTA* optimizer with a learning rate set to 0.1. Training via Stochastic Gradient Descent (SGD). Input batch size : 50, number of training iterations (epochs): 250. 10% of the training data for validation.

<sup>12</sup>features specific to a region in case of images or window of target sentences

of  $H$  embeddings to produce a local feature for the  $H$  target sentences. A local feature,  $c_k$  is generated from a window of embeddings  $RSV_{k:k+H-1}$  by applying a non-linear function (such as hyperbolic tangent) over the convoluted output. Mathematically,

$$c_k = f(W.RSV_{k:k+H-1} + b) \quad (3.7)$$

where  $b \in \mathbb{R}$  is the *bias* and  $f$  is the non-linear function. This operation is applied to each possible window of  $H$  target sentences to produce a feature map ( $c$ ) for the window size  $H$ .

$$c = [c_1, c_2, c_3, \dots, c_{N-H+1}] \quad (3.8)$$

A global feature is then obtained by applying *max-pooling* operation [169] over the feature map. The idea behind *max-pooling* is to capture the most important feature-one with the highest value-for each feature map. We describe the process by which one feature is extracted from one filter (red bordered portions in Figure 3.4 illustrate the case of  $H = 4$ ). The model uses multiple filters for each filter size to obtain multiple features representing the text. These features form the penultimate layer and are passed to a fully connected feed forward layer (with number of hidden units set to 100) followed by a *SoftMax* layer whose output is the probability distribution over the labels (*novel* or *non-novel*).

### 3.5.2 Results and Discussion

We proceed with the intuition: *redundancy* recognition would eventually lead us to *novelty* detection. The three datasets we use represent the most comprehensive resources that are publicly available and could be effectively used for document-level novelty detection.

#### On APWSJ

We take upon the results reported by [3] on APWSJ and use similar metrics to report the performance of our approach. Table 3.4 exhibits the effect of different redundancy measures (taken from [3]) on APWSJ considering both *absolutely redundant* and *somewhat redundant* documents as redundant or non-novel. Our RDV-CNN approach delivers comparable performance in a 10-fold cross-validation classification setting. Although the system suffers in identifying the *redundant* documents but succeeds to minimize the errors committed. It signifies the affinity of our architecture towards detecting *novel* documents. It is to be noted here that [3] conducted their experiments in an information filtering scenario using some thresholding scheme, where the **redundant or non-novel** documents were to be filtered from being delivered (*Not Deliv-*

### 3.5 Relative Document Vector-Convolutional Neural Network

ered). Only the *novel* documents were to be *Delivered* by the retrieval system. No learning was involved. Even APWSJ corpus was developed to support this Information Retrieval (IR) perspective of retrieving novel documents. Hence, there is a huge class imbalance. The presence of a relatively less number of *non-novel* documents in APWSJ (only 9.07%) and also that *some-what redundant* documents are considered as *redundant*<sup>13</sup> in our experiment may have hindered the learning of *redundant* patterns by our system. However, the superiority of our approach is established when we experiment with two other balanced datasets that closely approximates the semantic level redundancy we aim to capture.

Table 3.4: Results for Redundant class on APWSJ, *LM*  $\rightarrow$  Language Model, *Mistake*  $\rightarrow$  100-Accuracy. Except for RDV-CNN, all other numbers are taken from [3]

Measure	Recall	Precision	Mistake
Set Distance	0.52	0.44	43.5%
Cosine Distance	0.62	0.63	28.1%
LM:Shrinkage	0.80	0.45	44.3%
LM:Dirichlet Prior	0.76	0.47	42.4%
LM:Mixture Model	0.56	0.67	27.4%
<b>Proposed Approach (RDV-CNN)</b>	0.58	<b>0.76</b>	<b>22.9%</b>

#### On Webis-CPC-11

As stated earlier, we deem paraphrase detection as one close simulation of semantic level redundancy (non-novelty) detection. With this view we subject our model to detect passage-level paraphrase pairs from Webis-CPC-11. We investigate similar evaluation systems for novelty detection as carried out in [2]. The three novelty detection measures (Set Difference, Geometric Distance, Language Model), originally formulated by [3] and another based on *Inverse Document Frequency (IDF)* by [4] are our benchmarks for evaluation. Instead of setting a fixed threshold<sup>14</sup> as in these works we train a Logistic Regression (LR) classifier based on those measures to automatically determine the decision boundary.

**Baseline 1:** As a baseline we take *state-of-the-art* document embedding (*Paragraph Vector*) technique by [161] for document representation; concatenate the source and target document vectors and pass it to LR.

**Baseline 2:** Both the target and source sentence encodings (trained on SNLI) are passed to separate BiLSTM layers; the two resultant vectors are concatenated and passed to a Multi Layered Perceptron (MLP) with hidden dim of 2048 followed by classification via softmax.

Table 3.5 clearly shows that our RDV-CNN outperforms all the measures and baselines and maintains a significant supremacy to detect semantic-level redundant passages.

<sup>13</sup>as originally considered in [3]. We replicate the same experimental conditions for fair comparison.

<sup>14</sup>the weak thresholding algorithm reported in these works yield poor results

Table 3.5: Results for Paraphrase class on Webis-CPC (in %), IDF  $\rightarrow$  Inverse Document Frequency, LR  $\rightarrow$  Logistic Regression

Evaluation System	Description	P	R	$F_1$	A
Baseline 1	Paragraph Vector+LR	0.72	0.58	0.64	66.94%
Baseline 2	BiLSTM+MLP	0.71	0.73	0.72	70.91%
Novelty Measure 1 [3]	Set Difference + LR	0.71	0.52	0.60	64.75%
Novelty Measure 2 [3]	Geometric Distance + LR	0.69	0.75	0.71	70.23%
Novelty Measure 3 [3]	LM: Dirichlet Prior + LR	0.74	0.77	0.75	74.34%
Novelty Measure 4 [4]	IDF + LR	0.65	0.55	0.59	61.72%
<b>Proposed Approach</b>	<b>RDV-CNN</b>	<b>0.75</b>	<b>0.84</b>	<b>0.80</b>	<b>78.02%</b>

### On TAP-DLND 1.0

Next we experiment with our TAP-DLND 1.0 corpus. TAP-DLND 1.0 is well balanced and consists of a fair share of different levels (lexical as well as semantic) of textual views for both novel and non-novel documents. Since our objective here is to identify both novel and non-novel documents we report results for each class. We take the same baseline as we use for

Table 3.6: Results on TAP-DLND 1.0 , P  $\rightarrow$  Precision, R  $\rightarrow$  Recall, A  $\rightarrow$  Accuracy, R  $\rightarrow$  Recall, MLP  $\rightarrow$  Multi Layer Perceptron, N  $\rightarrow$  Novel, NN  $\rightarrow$  Non-Novel, IDF  $\rightarrow$  Inverse Document Frequency

Evaluation System	Description	P(N)	R(N)	$F_1(N)$	P(NN)	R(NN)	$F_1(NN)$	A
Baseline 1	Paragraph Vector+LR	0.75	0.75	0.75	0.69	0.69	0.69	72.81%
Baseline 2	BiLSTM+MLP	0.78	0.84	0.80	0.78	0.71	0.74	78.57%
Novelty Measure 1 [3]	Set Difference+LR	0.74	0.71	0.72	0.72	0.74	0.73	73.21%
Novelty Measure 2 [3]	Geometric Distance+LR	0.65	0.84	0.73	0.84	0.55	0.66	69.84%
Novelty Measure 3 [3]	LM:Dirichlet Prior+LR	0.73	0.74	0.74	0.74	0.72	0.73	73.62%
Novelty Measure 4 [4]	Novelty (IDF)+LR	0.52	0.92	0.66	0.66	0.16	0.25	54.26%
Feature Engineering [2]	Supervised features	0.77	0.82	0.79	0.80	0.76	0.78	79.27%
<b>Proposed Approach</b>	<b>RDV-CNN</b>	<b>0.86</b>	<b>0.87</b>	<b>0.86</b>	<b>0.84</b>	<b>0.83</b>	<b>0.83</b>	<b>84.53%</b>

Webis-CPC in the previous section. Our approach outperforms the popular semantic document embedding technique *Paragraph Vector* and another deep neural baseline (BiLSTM+MLP) by a considerable margin (12% and 6% in terms of accuracy, respectively). We also re-execute the other novelty detection measures by [3] (Set Difference, Geometric Distance, Language Model) and [4] (IDF) on TAP-DLND 1.0 and report the results. Our RDV-CNN even surpasses our feature-based earlier work [2] on TAP-DLND 1.0 by a considerable margin in identifying both novel and non-novel documents. This behavior we attribute to our customized architecture of mapping each target sentence to the nearest source sentence to produce a *relative* document representation being input to a CNN. The CNN then extracts the features from the *relative* document representation and finally classifies the incoming document. Results signify that our

### 3.5 Relative Document Vector-Convolutional Neural Network

---

network is able to learn the complex semantic interactions between source and target information necessary to conclude upon the state of novelty of a document. It is to be noted that except [2] and our baselines, all other measures that we consider for comparison were developed with an information retrieval perspective. The *p-values* for  $F_1$  score produced by 20 runs of our system against the baseline are less than 0.05 and hence the improvement is statistically significant and unlikely to be observed by chance in 95% confidence interval.

#### 3.5.3 Observations and Analysis

We scrutinize the results and arrive to the following observations:

- The RDV-CNN customized architecture of mapping each target sentence to the nearest source sentence for producing a *relative* document representation tackles the *relativity* criterion required for the problem. It facilitates CNN to extract the relevant features from the *relative* document representation which accounts for the better performance across the two datasets.
- Lexical approaches perform closer to our approach in detecting paraphrase pairs. This is due to the higher number of Named Entities (NEs) shared between those literary texts in Webis-CPC-11.
- Poor recall for non-novel class by IDF measure [4] is due to the existence of many new entity terms in the target documents for TAP-DLND 1.0.
- Lexical overlap based measures performs poorly in identifying *non-novel* documents in TAP-DLND 1.0. This behavior approves our discussion that we need semantic flair to address *non-novelty*. RDV-CNN bridges the gap.
- The deep baselines did not appear as promising to capture the semantic interactions between source and target sentences as did RDV-CNN.
- We thoroughly examined our gold and predicted class labels. Errors committed by our system are mostly due to presence of multiple source premise sentence for a target sentence. Our RDV-CNN could capture only one premise for a target sentence and map them. We

accommodate multiple premises in a later investigation.

- The universal sentence encoding generated by BiLSTM trained on SNLI are sometimes unable to capture the complex semantic interactions between source and target sentences. This mostly occurred for new NEs and out-of-vocabulary (OOV) words.

In this work, we show how we can leverage on *similarity* and *diversity* in a deep neural architecture for textual novelty detection. To the best of our knowledge, this work was the first to explore deep neural networks for document-level novelty detection.



## 3.6 Decomposable Attention for Novelty Detection

In our earlier work, we took a single source sentence as the premise for a target sentence. But that is a highly unlikely scenario. Usually, there could be multiple sources of information from which a piece of particular target information could be derived. Here in this work [66], we make use of neural attention mechanism to identify segments of texts in the source, which are more aligned with the target text and then jointly train them with a Multi-Layer Perceptron (MLP) to classify documents as novel or non-novel. We leverage the Decomposable Attention Model for Natural Language Inference [170] to work at the document-level for the problem. This approach yielded significant performance improvement of 2.9% on TAP-DLND 1.0 and 15.1% over APWSJ.

We arrive at a potentially feasible solution inspired by the phrase/sentence alignment problem in machine translation. We investigate which portion of texts in the source document(s) makes the target document appear *non-novel* to the reader and align the corresponding source-target text pairs to learn their interactions by a neural network. We hypothesize that for a *novel* document, there would be very little or no alignment with any portion of the source text. Eventually, we consider the sentence as the unit of information conveyance and look for their corresponding alignments in the respective texts (source and target). Let us consider the following *example* [2]:

- **d1:** *Singapore is an island city-state with a population of around 5.61 millions. Singapore's territory consists of one main island along with 62 other islets.*
- **d2:** *The Republic of Singapore is a sovereign country in Southeast Asia. The island city-state lies 137 kilometres north of the equator and has a dense population of approximately 5.6 million.*
- **d3:** *Singapore is a global commerce, finance and transport hub. Singapore has a tropical rainforest climate with no distinctive seasons, uniform temperature and pressure, high humidity, and abundant rainfall.*
- **d4:** *Singapore, an island city-state off southern Malaysia, lies one degree north of the equator. As of June 2017, the island's population stood at 5.61 million.*

It is fairly easy to conclude that document *d4* follows from *d1* and *d2*, by simply aligning the

two sentences in  $d_4$  with the first sentence of  $d_1$  and the second sentence of  $d_2$ . However, considering only  $d_3$  as the source, no such alignment is possible with  $d_4$ . Thus,  $d_4$  would be non-novel w.r.t  $d_1$  and  $d_2$  combined, but would appear novel if we consider  $d_3$  as the only source. Hence, we could say that once the reader goes through  $d_1$  and  $d_2$ , they could infer the contents of  $d_4$ . The inference is not exactly one-to-one as is the usual case in entailment literature. The example above describes a more likely multiple premise entailment [171] scenario. Here the premise of the first target sentence in  $d_4$  is both the first sentence of  $d_1$  and the second sentence of  $d_2$ . However, if the reader goes through only  $d_3$ , then document  $d_4$  would seem to contain new information to them. Quite interestingly here we could see a relation brewing between text alignment and sentence inference while judging the novelty of a piece of text. To effectively model this relation, we need efficient sentence representations. Hence we train our sentence representations on a large scale natural language inference dataset (the Stanford Natural Language Inference (SNLI) corpus) to capture the essence of the inference knowledge in our sentence embeddings. As discussed earlier, we manifest the alignment perspective via the neural attention mechanism as substantiated in our further discussions. We set out to investigate this idea and see whether a neural network could learn text alignment and correctly identify a document as novel or redundant. We achieve this alignment via the attention mechanism [172], popular in deep neural networks. Here novel document  $d_3$  is relevant, yet diverse w.r.t others. Word/Phrase level alignment is handled with reasonable accuracy via attention in entailment literature [170]. We leverage their idea for sentence-level entailment to work in our case for ascertaining novelty of documents; a kind of transfer learning (textual entailment to novelty detection) approach to the problem concerned. One key challenge in this work is to generate a sentence representation that could effectively model this alignment perspective. We do so via the *inner-attention* based sentence encoder [173] trained on the very large and semantically rich SNLI corpus [166].

### 3.6.1 Proposed Method

As discussed earlier, we intend to make use of attention mechanism to identify the potential contributing sections in the source document(s) that makes a target document appear *non-novel* to the reader. We draw inspiration from the work of [170] on decomposable attention for natural language inference. Using *Attend*, *Compare* and *Aggregate* steps in the model, they successfully show the importance of attention to identify the contributing sections in a sentence for inference decisions. Natural Language Inference is one such task that closely resembles the notion of *non-novelty*. The relation of novelty detection with entailment is extensively studied in the RTE-TAC novelty subtasks and associated literature [93]. This is also why we are motivated

### 3.6 Decomposable Attention for Novelty Detection

---

to base our model on the knowledge learned from an entailment dataset. Hence, we proceed to transfer inference knowledge from an external NLI dataset to our text representations, train our proposed model on the novelty detection datasets, and thereby investigate the performance. We hypothesize that for a piece of *novel* text there should be no contributing section (i.e., text having similar/overlapping information content: either lexically or semantically) in the source texts. We leverage the word-level attention model to sentence-level attention to effectively model document-level novelty detection. Instead of generating a complex joint representation of the source and target texts with dominant deep neural paradigms, the decomposable attention model [170] relies on simple *alignment* of target text with the source text. The decomposable model is also found to be efficient in terms of the number of parameters as compared to the other models for modelling entailment pairs.

Initially, we encode our sentences using intra-sentence attention trained on the SNLI corpus to capture the rich semantic perspectives involved in sentence level inference decisions. Then we make use of the decomposable attention model to learn the notion of document-level novelty and redundancy from the aggregated representation. We depict the overall architecture in Figure 3.5.

Here  $T_1$  is the *target* document whose state of *novelty* is to be determined against the source document(s)  $S_1, S_2, S_3$ . Although we investigate novelty at the document-level, we rely on sentence-level interactions among the source and target texts. Hence we split the source and target texts into component sentences and generate the sentence encodings. Please note that our architecture is not end-to-end trained. In the first phase, we train the sentence encodings on the semantically rich SNLI dataset. In the second phase, we train the decomposable attention model on the novelty detection datasets (APWSJ, TAP-DLND 1.0). The sentences in the novelty datasets are vectorized with the sentence encodings from SNLI training.

#### Sentence Encoding

The task of *Novelty Detection* requires high-level understanding and reasoning about semantic relationships within texts. Textual Entailment (TE) or Natural Language Inference (NLI) is one such task which exhibits such complex semantic interactions. Hence, we train our sentence encodings on the very large (about 570k text pairs) and semantically rich Stanford Natural Language Inference (SNLI) dataset [166]. We use open source GloVe vectors trained on Common Crawl 840B with 300 dimensions as fixed word embeddings. There are several other state-of-the-art sentence embedding techniques like the Universal Sentence Encoder [174], self-attentive

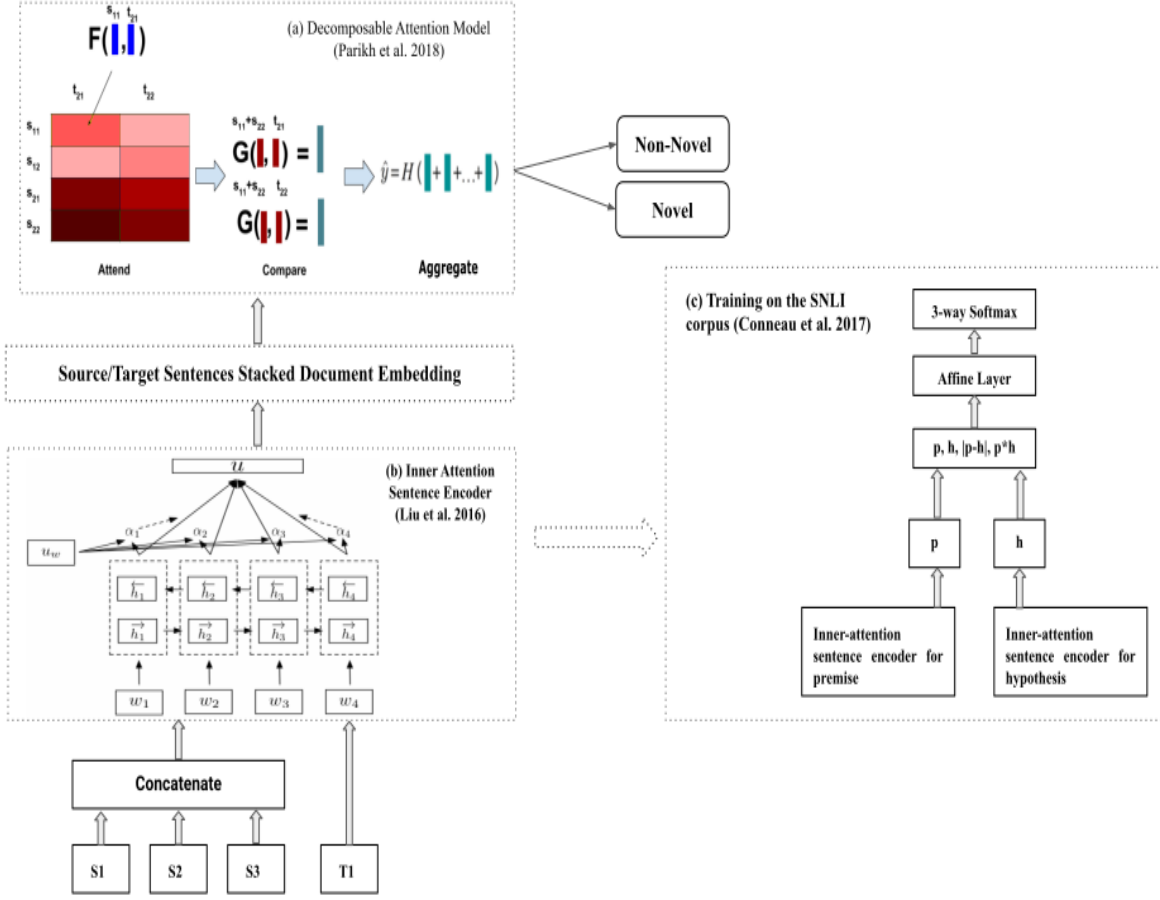


Figure 3.5: Overall architecture for document-level novelty detection. Component (b) is the inner-attention sentence encoder. Component (c) shows how the inner attention sentence encoder is trained on the SNLI corpus. Component (a) is the sentence-level decomposable attention model we use in our work for document level novelty detection.  $s_{11}, s_{12}$  represents the two sentences in source document  $d_1$  in the example introduced in Section 3.6.  $s_{21}, s_{22}$  are the two sentences in source document  $d_2$ .  $d_1$  and  $d_2$  are concatenated to form a single source document.  $t_{21}, t_{22}$  are the two sentences in target document  $d_4$ . Simply reading the example we can conclude that  $t_{21}$  and  $t_{22}$  directly follow from  $s_{11}$  and  $s_{22}$ .  $d_4$  is redundant if we consider  $d_1$  and  $d_2$  as the source documents.

sentence encoder [175], unsupervised sentence embedding using weighted average of word vectors [176], etc., but we use the SNLI-trained sentence encoder [6] in order to have the NLI semantic interactions in our sentence embeddings.

### Stanford Natural Language Inference Dataset

The SNLI corpus [166] is a collection of 570k human-written English sentence pairs manually labelled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE). As discussed earlier, we seek to transfer the inference knowledge of SNLI to our

### 3.6 Decomposable Attention for Novelty Detection

---

sentence embeddings.

#### Training the sentence encoder on external linguistic resource: SNLI

We already explained the reason why we chose SNLI to generate our sentence embeddings in preceding sections. Authors in [6] demonstrated that sentence encoder trained on natural language inference corpus could learn sentence representations that capture universally useful features. We follow the idea of a Siamese Neural Network (Figure 3.5(c)). It denotes that two identical sentence encoders share the same set of weights during training, and the two sentence representations (premise  $p$  and the hypothesis  $h$ ) are then combined to generate a relative vector for classification. A typical architecture of this kind uses a shared sentence encoder that outputs a representation for the premise  $p$  and the hypothesis  $h$ . Three matching methods [177] are applied to extract the relations between  $p$  and  $h$ : (i) concatenation of the two representations  $(p, h)$ ; (ii) element-wise product  $p * h$ ; and (iii) absolute element-wise difference  $|p - h|$ . The resulting vector, which captures information from both premise and the hypothesis, is fed into a 3-class classifier (neutral, entailment, contradiction: class labels in SNLI) consisting of multiple fully-connected layers culminating in a softmax layer.

#### Bi-LSTM + Inner-Attention Sentence Encoder

We use an inner attention sentence encoder [173] to generate a representation  $u$  of an input sentence. This encoder employs attention mechanism on the representation produced in previous hidden state of a Bi-directional LSTM to attend important words in the sentence itself. This is inspired from the *inner-attention* idea from [173] where the author says that readers usually form a rough intuition about which part of the sentence is more important, usually from their past experiences. For a sequence of  $T$  words  $\{w_t\}_{t=1, \dots, T}$ , a Bi-LSTM computes a set of  $T$  vectors  $\{h_t\}_t$ . For  $t \in [1, \dots, T]$ ,  $h_t$  is the concatenation of a forward LSTM and a backward LSTM that read the sentences in two opposite directions:

$$\begin{cases} \vec{h}_t = \overrightarrow{LSTM}_t(w_1, \dots, w_T) \end{cases} \quad (3.9)$$

$$\begin{cases} \overleftarrow{h}_t = \overleftarrow{LSTM}_t(w_1, \dots, w_T) \end{cases} \quad (3.10)$$

$$\begin{cases} h_t = [\vec{h}_t, \overleftarrow{h}_t] \end{cases} \quad (3.11)$$

The attention mechanism is formalized as :

$$\bar{h}_i = \tanh(Wh_i + b_w) \quad (3.12)$$

$$\alpha_i = \frac{e^{\bar{h}_i^T u_w}}{\sum_i e^{\bar{h}_i^T u_w}} \quad u = \sum_i \alpha_i h_i \quad (3.13)$$

where  $(h_1, \dots, h_T)$  are the output hidden vectors of a Bi-LSTM. These are fed to an affine transformation  $(W, b_w)$  which outputs a set of keys  $(\bar{h}_1, \dots, \bar{h}_T)$ . The  $\alpha_i$  represents the score of similarity between the keys and a learned context query vector  $u_w$ . These weights are used to produce the final representation  $u$ , which is a weighted linear combination of the hidden vectors.

### The Decomposable Attention Model

The decomposable attention model was first proposed by [170] for word-level alignment to model sentence-level inference. We take inspiration from their work and make use of decomposable attention for sentence alignment to model identification of document-level novelty. The objective of the decomposable model is to decompose the task into the subproblems that are solved separately. In contrast to recent dominant complex approaches for natural language inference this approach only relies on the alignment of local text substructure to aggregate to a prediction.

Let  $x = [x_1, x_2, \dots, x_m]$  denote the set of all sentences concatenated from the source documents and  $y = [y_1, y_2, \dots, y_n]$  denote the set of sentences in the target document. The corresponding sentence vector representation of these sentences are denoted by  $\bar{x} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m]$  and  $\bar{y} = [\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n]$ . The core decomposable model (Figure 3.5(a)) consists of the following three components, which are trained jointly:

1. **Attend:** The attention layer uses a variant of neural attention proposed in ([172]). We implement it using a feed-forward neural network  $F$  that is applied to both sentences separately. We soft align the elements of  $\bar{x}$  and  $\bar{y}$  via attention mechanism and decompose the problem into the comparison of aligned sentences. We obtain the unnormalized attention weights  $e_{ij}$ , computed by a function  $F'$ , which decomposes as:

$$e_{ij} := F'(\bar{x}_i, \bar{y}_j) := F(\bar{x}_i)^T F(\bar{y}_j) \quad (3.14)$$

$F$  is a feed-forward neural network with ReLU activations. These attention weights are normalized as follows :

$$\begin{aligned} \beta_i &= \sum_{j=1}^n \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{ik})} \bar{y}_j \\ \alpha_j &= \sum_{i=1}^m \frac{\exp(e_{ij})}{\sum_{k=1}^m \exp(e_{kj})} \bar{x}_i \end{aligned} \quad (3.15)$$

Here,  $\beta_i$  is the text segment in  $\bar{y}$  that is (softly) aligned to  $\bar{x}_i$  and the vice versa for  $\alpha_j$ .

### 3.6 Decomposable Attention for Novelty Detection

---

2. **Compare:** Next step is to compare the soft-aligned document matrices. Similarly to the previous step, we use a feed-forward neural network  $G$  to compare the aligned text segments  $\{(\bar{x}_i, \beta_i)\}_{i=1}^m$  and  $\{(\bar{y}_j, \alpha_j)\}_{j=1}^n$

$$\mathbf{v}_{1,i} = G([\bar{x}_i, \beta_i]) \quad \mathbf{v}_{2,j} = G([\bar{y}_j, \alpha_j]) \quad (3.16)$$

where the brackets  $[\cdot, \cdot]$  denote concatenation.

3. **Aggregate:** The last part of the core model architecture is the aggregation layer. All this layer does is a column-wise sum over the output of the comparison network:

$$\mathbf{v}_1 = \sum_{i=1}^m \mathbf{v}_{1,i} \quad \mathbf{v}_2 = \sum_{j=1}^n \mathbf{v}_{2,j} \quad (3.17)$$

and feed the output through a final classifier  $H$ , which is again a feed forward network followed by a linear layer:

$$\hat{y} = H([\mathbf{v}_1, \mathbf{v}_2]) \quad (3.18)$$

where  $\hat{y}$  represents the predicted scores for each class (novel or non-novel). Output layer of the classifier is softmax normalized so that we obtain a probability distribution over target classes and consequently the predicted class is given by  $\hat{y} = \text{argmax}_i \hat{y}_i$

#### Hyperparameter Settings and Parameter tuning

The training of the sentence encoder on SNLI is carried out with 20 epochs, batch size=64 with SGD optimizer and learning rate set to 0.001. The output sentence vector dimension  $SENT\_DIM$  is 2048. For the decomposable model we use epochs = 30, batch size = 25, Adam optimizer with learning rate set to 0.001 and the loss function as categorical cross entropy. The comparison layer compares each target sentence with its corresponding alignment, finally outputs a vector of dimension  $2*SENT\_DIM$ . Next, the entailment layer is a fully connected layer with two hidden layers and  $SENT\_DIM$  neurons with RELU activation. Finally, softmax layer is used for the classification.

The Bi-LSTM inner attention sentence encoder has two affine layers without bias. The first and second layer in the SNLI training has 512 neurons each and the classification layer has three neurons (entailed, contradiction, neutral). The attention layer in the decomposable model

is actually a Bi-RNN over the sentence vector (2048 dimension) and then dot product. The compare step is a Bi-RNN over concatenation of original and the aligned sentence. Finally, in the aggregate layer, we have an affine transformation to 2048 neurons and finally a prediction layer of two neurons.

### 3.6.2 Evaluation

We evaluate our proposed approach on two benchmark datasets (APWSJ and TAP-DLND 1.0) and present the results in the following sections. As discussed before, we have two segments in our proposed model:

1. The inner attention sentence encoder trained on the very large scale SNLI corpus.
2. The decomposable attention model trained on the document-level novelty detection datasets separately with the proposed method.

Thus our model is not end-to-end trained. We leverage the knowledge of natural language inference (first segment) to classify documents as novel or non-novel (core task; second segment).

### Baselines and Comparing Systems

We adopt the following baselines to investigate the strength of various factors in our model as well as help in ablation studies. We also compare with our earlier approaches to better scrutinize our performance.

- **Baseline 1 (Without Decomposable Attention Model):** We choose this baseline to see the effect of withholding the second segment (i.e., the decomposable attention module) of our model. Here, we train the sentence encoder (based on inner attention) on SNLI. We generate the document matrix for the source<sup>15</sup> and target documents by stacking the corresponding sentence encodings. A Bi-LSTM layer encodes each document matrix and produces a fixed-sized vector of dimension *SENT\_DIM* for the source and target document(s) separately. The two resultant vectors are then concatenated and passed to a fully connected entailment layer with two hidden layers having *SENT\_DIM* neurons (basically a Multi-Layered Perceptron) and *RELU* activation followed by classification via softmax.
- **Baseline 2 (Without Inner Attention Sentence Encoder):** Instead of using attention mechanism over the hidden states of the Bi-LSTM to generate a representation  $u$  of an input sentence, we select the maximum value over each dimension of the hidden units

---

<sup>15</sup>concatenated for multiple source documents



### 3.6 Decomposable Attention for Novelty Detection

---

(max pooling) [6].

- **Baseline 3 (Without Pre-training on SNLI):** To see how much pre-training of the sentence encoder with the large-scale SNLI affects our system performance, we ablate the training of the sentence encoder on SNLI. We use the pre-trained paragraph vector<sup>16</sup> (*doc2vec*; [161]) to generate the sentence encodings. Here we want to see how much of the contribution of the pre-training step comes from training on SNLI as opposed to simply using pre-trained embeddings (*doc2vec*-Distributed Bag Of Words) on a news dataset (Associated Press News). We then train the decomposable model, and finally classify the target document.
- **Comparing Systems (Previously Published Results):** We compare the performance of our proposed system against the novelty measures proposed by [3] for APWSJ. The three novelty detection measures from [3] are Set Difference (*Novelty Measure 1*), Geometric Distance (*Novelty Measure 2*), and Language Model (*Novelty Measure 3*). For TAP-DLND 1.0, we re-implement these measures along with another one (*Novelty Measure 4*) which is based on inverse document frequency [4]. For comparison, we also consider another approach based on calculating the relative entropy of a document [178]. Instead of setting a fixed threshold<sup>17</sup> as in these works we train a Logistic Regression (LR) classifier based on those measures to automatically determine the decision boundary.

We also compare the current approach with our earlier attempts on TAP-DLND 1.0: feature-based [2] and the RDV-CNN architecture [64]. In our feature-based solution [2] to the problem, we make use of several features like document similarity, divergence, named-entities, keywords, etc. In our subsequent work [64], we employ a CNN based deep model that learns the notion of novelty and redundancy from a derived vector representation from source and target documents which we term as the Relative Document Vector (RDV). We trained the sentence embeddings on the SNLI dataset using a BiDirectional LSTM with max pooling [6] technique. We then pulled the nearest source sentence to a given target sentence and concatenated them using the representation of [177]. Finally we stacked the source encapsulated target sentence representations to form the Relative Document Vector. We used a CNN to extract features from the RDV to classify a given target documents as novel or non-novel.

---

<sup>16</sup><https://github.com/jhlau/doc2vec>

<sup>17</sup>the weak thresholding algorithm reported in these works yield poor results

Table 3.7: Results on TAP-DLND 1.0, P→ Precision, R→ Recall, A→ Accuracy,  $F_1$  → Average F-Score, N→ Novel, NN→ Non-Novel, \* →measures from [3] with Logistic Regression (LR), ‡ →measure from [4] with Logistic Regression (LR), † →10-*fold* cross-validation output, IDF→Inverse Document Frequency

System	P(N)	R(N)	P(NN)	R(NN)	$F_1$	A(%)
Baseline 1†	0.74	0.85	0.88	0.78	0.81	81.4
Baseline 2†	0.74	0.46	0.69	0.88	0.69	70.4
Baseline 3†	0.74	0.43	0.68	0.89	0.68	69.5
Novelty Measure 1*	0.74	0.71	0.72	0.74	0.72	73.2
Novelty Measure 2*	0.65	0.84	0.84	0.55	0.69	69.8
Novelty Measure 3*	0.73	0.74	0.74	0.72	0.73	73.6
Novelty Measure 4‡	0.52	0.92	0.66	0.16	0.45	54.2
Dasgupta and Dey (2016) [178]	0.63	0.72	0.77	0.66	0.69	68.2
Feature-based approach [2]†	0.77	0.82	0.80	0.76	0.78	79.3
RDV-CNN [64]†	0.86	0.87	0.84	0.83	0.85	84.5
<b>Proposed Approach†</b>	0.85	0.85	<b>0.89</b>	<b>0.89</b>	<b>0.87</b>	<b>87.4</b>

It is to be noted here that except [2], [64], [178], and our baselines; all the other comparing systems were developed from an information retrieval perspective. Whereas in our earlier attempt [64] we use a CNN based model on a derived source-target representation, here in our current work, we use an attention-based model to aggregate aligned information from source and target documents to identify novelty and redundancy.

### 3.6.3 Results and Discussions

We show the comparatively better performance of our approach in the evaluation figures of Table 3.7 (TAP-DLND 1.0) and Table 3.8 (APWSJ). The current approach outperforms our previous two *state-of-the-art* approaches by a margin of 8.1% and 2.9% in terms of accuracy. Our initial system is feature-based [2]. Although our second system is a deep neural architecture [64], the current approach can surpass its performance with lesser complexity in terms of the number of parameters used. Our baselines also serve as a good means of ablation study.

- We observe that the novelty measures [3, 4] do not perform well in identifying novel or non-

### 3.6 Decomposable Attention for Novelty Detection

Table 3.8: Results for Redundant/Non-Novel (NN) class on APWSJ,  $LM \rightarrow$  Language Model,  $Mistake \rightarrow 100$ -Accuracy,  $\dagger \rightarrow 10$ -fold cross-validation output,  $*$   $\rightarrow$  results taken from [3]

Measure	R(NN)	P(NN)	Mistake (%)
Set Distance*	0.52	0.44	43.5
Cosine Distance*	0.62	0.63	28.1
LM: Shrinkage*	0.80	0.45	44.3
LM: Dirichlet Prior*	0.76	0.47	42.4
LM: Mixture Model*	0.56	0.67	27.4
RDV-CNN [64] $\dagger$	0.58	0.76	22.9
<b>Proposed Approach<math>\dagger</math></b>	<b>0.86</b>	<b>0.92</b>	<b>7.8</b>

novel documents. This is because these measures do not consider the semantics involved in recognizing novelty. These lexical measures also do not consider context information which is of utmost importance to operating at the discourse level.

- Without the inner attention sentence encoder (Baseline 2) we see that the system does not perform good to classify the target documents. This resonates the observation made by [173] that it is necessary to emphasize certain sections of texts (achieved via inner attention) based on the already seen data or neighbourhood data (i.e., the context).
- **Impact of inference knowledge from SNLI:** When we remove the SNLI pre-training of the sentence encoder (Baseline 3), the performance of the proposed system drops by a considerable amount ( $\sim 18\%$  in terms of accuracy). This behaviour supports our intuition that upon training on the large-scale SNLI corpus, the sentence embeddings can capture the semantics involved in understanding NLI. As discussed in the preceding sections, NLI is one such phenomenon that closely simulates our reckoning of non-novelty at the semantic level. Thus we could say that the inference knowledge transfer from SNLI highly correlates with the text alignment perspective for novelty detection.
- Baseline 1 shows the effect of not having the decomposable attention module in our system. Although we have the inner attention sentence vectors pre-trained on SNLI, it is evident

that it costs us  $F_1$  by a margin of  $\sim 6$  points. This is because, in this model, the resultant vectors obtained from the source and target document(s) do not have the information: which source text should a target sentence be attending to determine the novelty of the target sentence. We deem that this step is essential to adjudge whether a document is novel/non-novel w.r.t. a set of given documents (or known information). Eventually, the novelty of constituent sentences would lead us to the judgment of novelty of the given document. Particularly we see that the recall for Non-Novel (NN) documents is low which points that the pairing of target sentence with its corresponding source sentence (which we achieve via attention) is important to create the resultant vector before feature extraction and classification via a neural network.

- Our current approach supersedes our previous hand-crafted feature based system on TAP-DLND 1.0 [2] by a margin of  $\sim 8\%$  in terms of accuracy. This is particularly encouraging in the sense that our proposed model can capture the textual characteristics required to understand document novelty from the data itself. Also, our current proposed approach shows improvement over our recently proposed RDV-CNN [64] architecture for document-level novelty detection ( $\sim 3\%$  in terms of accuracy). The current attention based architecture is also simple as compared to RDV-CNN in terms of implementation and in the order of parameters used.
- Table 3.8 is a strong testimony that the proposed approach is effective to identify the non-novel documents. The alignment of the target sentence to the corresponding source sentence(s) and subsequent feature discovery via neural layers prove to be appropriate for recognizing redundant documents in APWSJ. It is particularly encouraging to see that even though there are a lesser number of non-novel documents in APWSJ (which is more likely a practical scenario), attention mechanism enables to identify the sentences which actually contribute to making a target sentence *non-novel*.
- The results strongly corroborate our intuition that textual alignment of sentences (between the source and target documents) could lead to a better understanding of document novelty. The alignment should not necessarily be on a one-to-one sentence basis. Rather a target sentence may have multiple information sources (sentences). The multiple premise cases are well handled with the distributed attention weights over multiple sentences, as discussed in the Attend step of the decomposable attention model. But for efficient attending of

### 3.6 Decomposable Attention for Novelty Detection

source document sentences (identifying the premise) by a target document sentence (the hypothesis), we need sentence embeddings that could manifest the inference perspective. We achieve that via our inner-attention sentence encoder trained on the large-scale SNLI corpus. Thus a relation between textual inference and alignment is established, which proves crucial to novelty detection.

The accuracy of the proposed model is statistically significant over the baselines (two-tailed t-test,  $p < 0.05$ ).

#### 3.6.4 Analysis

The strength of our approach stems from the very objective we start to investigate: *alignment* via *attention* with the value addition of *inference*. Considering the example in section 3.6, we employ our approach and plot the attention weights obtained from the *Attend* step in Figure 3.6. From the example we consider  $d_4$  as the target document against the source documents  $d_1$  and  $d_2$ . We concatenate  $d_1$  and  $d_2$  to form one single source having four sentences. From the attention weights we can clearly see that the two target sentences (in  $d_4$ ) are highly attending the first and fourth source sentences (the corresponding attention weights are high), signifying a multiple premise scenario. However, the second and third source sentence having diverse information finds less alignment or attention of the target sentences. Our method correctly predicts the

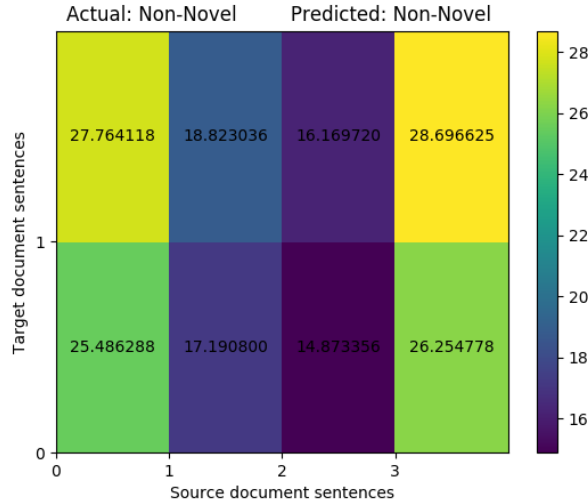


Figure 3.6: Attention matrix visualization via heat map for the example in Section 3.6.  $d_1$  and  $d_2$  are concatenated to form the source document.  $d_4$  is the target document.

class label, as well. Thus, if there is less attention (less attention weights for most sentence pairs), which means fewer appropriately aligned sentences, which in turn signifies the majority of the sentences in the target are distant from each sentence in the source: the tendency of the

target document is towards novelty. The target document sentences are not highly attending or aligned to any source document sentence, or the target sentences found no premise in the source document(s); the target document supposedly has new information. The corresponding heat map would look like Figure 3.7(a) signifying that the document is *novel*. Figure 3.7(b) is the heatmap of a correctly predicted non-novel document from TAP-DLND 1.0 pitched against the corresponding source document(s).

### 3.6.5 Error Analysis

We perform a thorough analysis of our predictions. Figure 3.7 (c,d,e) show some error instances. Majority of the errors committed by our system are due to the presence of:

1. Multiple complex premise sentences in the source documents for a target sentence. The attention weights are evenly distributed, sometimes making it hard to identify the exact premises.
2. Long compound sentences and conveying a greater amount of information, resulting in misalignment.
3. Higher number of Named-Entities, which are often over-emphasized by the attention model. This often contributes to false negatives. In spite of conveying different information but due to the presence of same named-entities, sentences are misaligned (given higher attention weights).
4. Dichotomy in annotation judgments in both the datasets. We manually went through the annotation judgments and in certain cases were not sure about the ground-truth. Subjectivity plays a crucial role here. The novelty appetite is not the same for all readers [13]. The amount of new information that makes a document appear novel to one reader may not be the same to another reader.
5. We cannot always establish a simple mapping between sentences in the source and target documents. Sometimes target documents consist of background information (world knowledge; pragmatics) that has relevance with the source but are not explicitly mentioned.
6. Errors in sentence splitting and the difference in document lengths. Some source documents after concatenation are too long and manifest all the information in the target document. Hence the target document should be non-novel. But the information in the concatenated source document is distributed all across the source sentences. Thus there is no emphasis on one or two source sentences. Hence finding a suitable source text segment to get aligned

### 3.6 Decomposable Attention for Novelty Detection

---

with a target sentence is difficult. Therefore although the target is non-novel, but the model predicted it as a novel.

We pulled out some examples from TAP-DLND 1.0 and demonstrate the error categories (See heatmaps in Figure 3.7 (c,d,e)). The following instances are actually Non-Novel but predicted as Novel.

- **Instance #45:** Observations (Figure 3.7c)
  - Target sentence #15 has many premises so its attention values are spread throughout the source documents decreasing its importance.
  - Target sentence #11 conveys important information but its complex structure and long length made it difficult for the model to capture its attention values.
  - Target sentence #6 has some named entities (NEs) which unusually increased its attention values.
- **Instance #381:** Observations (Figure 3.7d)
  - Target sentence #2 conveys important information but due to its long length resulted in low attention values
  - Target sentence #5 is over emphasized due to high NEs
- **Instance #495:** Observations (Figure 3.7e)
  - Target sentence #0 and sentence #2 are important but due to their long length and complexity, the model could not capture their attention values appropriately.

We analyzed 250 false positive and false negative cases. Category 1 and Category 2 errors together contributed to nearly ~71% of the misclassified instances.

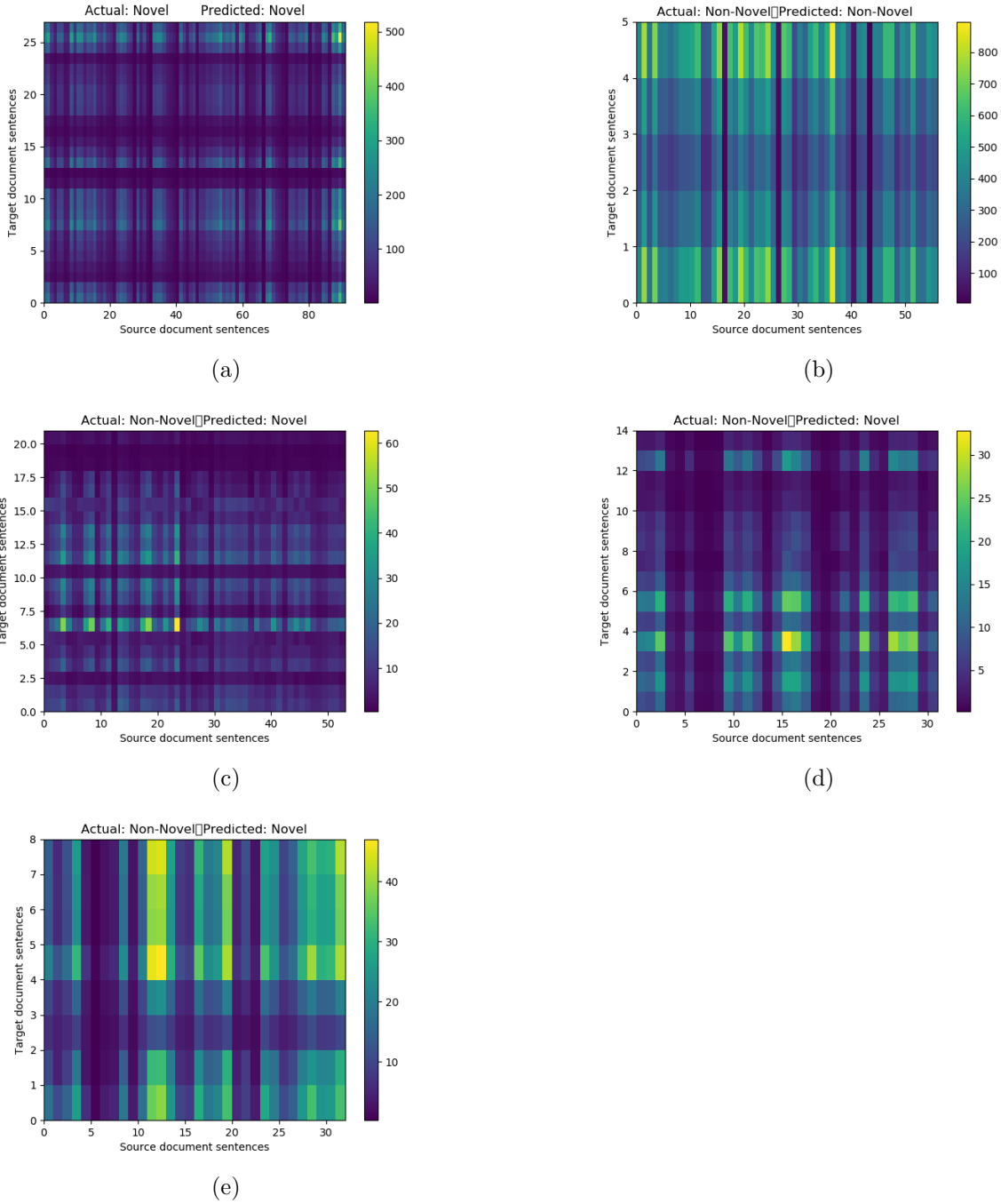


Figure 3.7: Attention matrix visualization via heat map for a correctly predicted (a) novel and (b) non-novel document from TAP-DLND 1.0. (a) Many dark patches signify that most of the target sentences are not highly attending to any source sentence. Hence, may contain sufficient new information. (b) Lesser dark patches indicate that the target sentences are highly aligned to the source sentences and may contain redundant information. Wrongly predicted instances  $\rightarrow$  (c), (d), (e).



## 3.7 Quantifying Novelty

In this work, we try to quantify the amount of new information. The appetite of novel information is different with different readers [13] which makes the problem a very subjective one. Hence figuring out the amount of new information in a document and then letting the users decide the category of the document (novel, non-novel, partially novel) appears an interesting direction to investigate into this problem. We study the problem: *to predict the novelty score of a document based on the document(s) or information already seen by the system*. We carry our work on the TAP-DLND 2.0 dataset described earlier.

### 3.7.1 Problem Definition

We define the problem as associating a qualitative novelty score to a document based on the amount of new information contained in it. Let us consider the following example:

**Source Text:** *Singapore, an island city-state off southern Malaysia, is a global financial center with a tropical climate and multicultural population. Its colonial core centers on the Padang, a cricket field since the 1830s and now flanked by grand buildings such as City Hall, with its 18 Corinthian columns. In Singapore's circa-1820 Chinatown stands the red-and-gold Buddha Tooth Relic Temple, said to house one of Buddha's teeth.*

**Target Text:** *Singapore is a city-state in Southeast Asia. Founded as a British trading colony in 1819, since independence it has become one of the world's most prosperous, tax-friendly countries and boasts the world's busiest port. With a population size of over 5.5 million people it is a very crowded city, second only to Monaco as the world's most densely populated country.*

The task is to find the novelty score of the target text w.r.t the source text. It is quite clear that the target text is having new information with respect to the source except that the first sentence in the target contains some redundant content (*Singapore is a city-state*). Analysing the first sentence in the target text we get two information: that *Singapore is a city-state* and *Singapore lies in Southeast Asia*. Keeping the source text in mind, we understand that the first part is *redundant* whereas the second part has new information, i.e., we can infer that 50% information is novel in the first target sentence. Here, we consider only the surface-level information in the text and do not take into account any pragmatic knowledge of the reader regarding the geographical location of Singapore and Malaysia in Asia. Here, our new information appetite is more fine-grained and objective.

Now let us attach a qualitative score to each of the three target sentences as 0.5, 1.0, 1.0, signifying 50% new information (0.5) and total new information (1.0), respectively. The cumulative sum comes to 2.5 which says that the target text has 83.33% new information w.r.t the source text<sup>18</sup>. This scoring mechanism, although straightforward, intuitively resembles the human-level perception of the amount of new information. However, we do agree that this approach attaches equal weights to long and short sentences. Long sentences would naturally contain more information whereas short sentences would convey less information. Also, we do not consider the relative importance of sentences within the documents. However, for the sake of initial investigation and ease of annotation<sup>19</sup>, we proceed with this simple quantitative view of novelty and create a dataset that would be a suitable testbed for our experiments to predict the document-level novelty score.

### 3.7.2 Methodology

Our main intention is to predict the novelty score of a document given a set of relevant documents already seen by the model. Having created an appropriate dataset, we design an architecture, which encodes the target document information jointly with the source information in one single unit and then makes use of a deep Convolutional Neural Network (CNN) to predict the novelty score. This jointly encoded target document representation is particularly important here, because, for novelty, the context of the topic in concern plays a pivotal role.

#### Premise Selection

Selecting (Recalling) the appropriate source document(s) is essential for novelty search [179]. We relate this with the *Two-Stage theory* of human recall [180] consisting of **Phase-I: Search and Retrieval** and **Phase-II: Recognition**. Here to realise Phase-I from the pool of all source documents (simulating the memory or information already known), we select the top 10 documents which could be the potential source of a given target document. We take Named-Entities similarity [181, 182] to retrieve the 10 most similar documents. The Recall@10 here at this stage is 0.93. For Phase-II, out of the retrieved 10, we further filter three potential source documents<sup>20</sup> via the Word Mover’s Distance [183]<sup>21</sup>. Less is the distance; higher is the ranking for relevance. The Recall@3 at this stage is 0.94.

---

<sup>18</sup>if all the sentences were tagged as novel, the score would have been 3.0 indicating 100% novel information in the target text

<sup>19</sup>identifying and annotating an information unit would have been complex. However, we plan for further research with annotation at the phrase-level and with relative importance scores

<sup>20</sup>we know that each event has three source documents in the corpus

<sup>21</sup>WMD works reasonably well for similarity search in the semantic space for shorter documents

### 3.7 Quantifying Novelty

#### Source Encapsulated Target Document Vector (SETDV)

As discussed in [2], the novelty of texts is to be always determined with respect to a set of relevant information already known about the topic in concern. Thus, one cannot ascertain the novelty of a document unless s/he sees relevant source/prior information. However, the first document about a topic would always be novel if we consider a topical document stream. Novel information is often argued as an update over the existing knowledge [2]. Hence to capture this perspective, we create a target document representation that jointly encodes the target and relevant source information, which we term as the **Source Encapsulated Target Document Vector (SETDV)**. The idea is simple: we pull out the nearest source sentence corresponding to a target one and encapsulate them in one single representative sentential unit. Figure 3.8 shows the SETDV-CNN architecture. Here,  $T_1$  is the *target* document whose *novelty* score is to be determined against the source document(s)  $S_1, S_2, \dots, S_M$  i.e., to say the objective is to automatically figure out the novel information content in  $T_1$ , once the machine has already seen/scanned  $S_1, S_2, \dots, S_M$ .

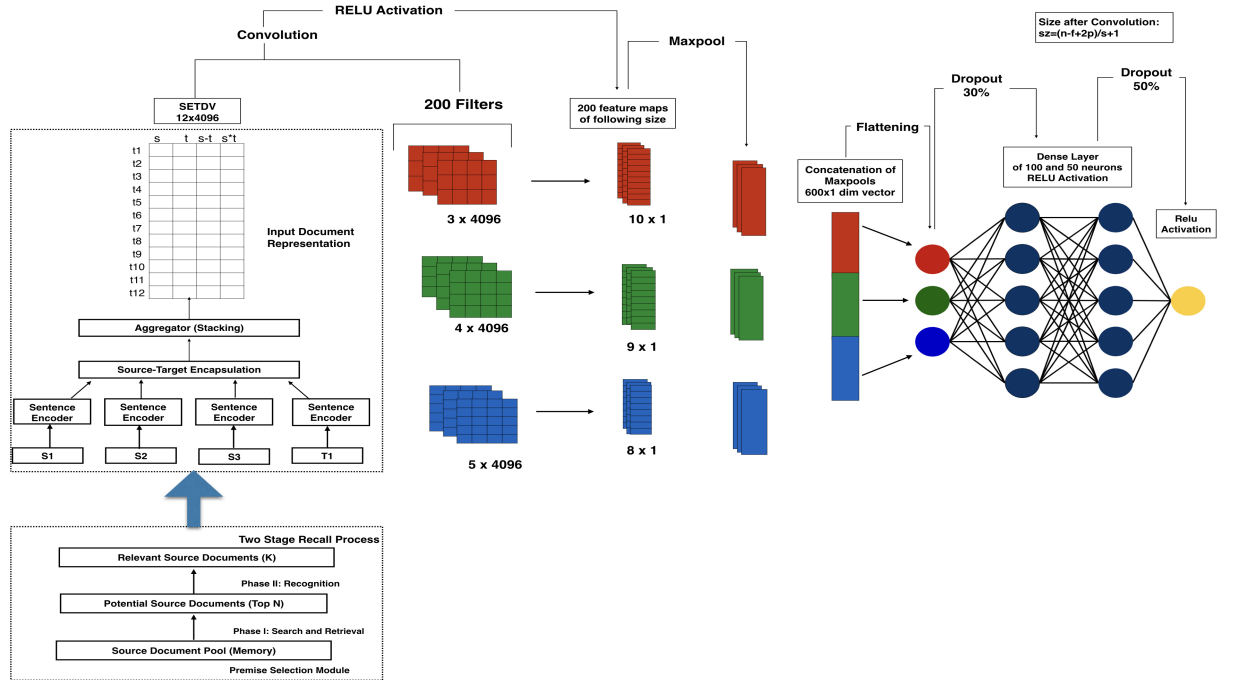


Figure 3.8: The overall Novelty Score Prediction Architecture (SETDV-CNN) with 12 sentences in the target document ( $T_1$ ) and  $S_1, S_2, S_3 \rightarrow$  source documents

#### Sentence Encoder

Instead of encoding the entire document, we encode the sentences. This makes sense as we even annotate at the sentence level. Following from [6], we train a sentence encoder on the semantically

rich large-scale Stanford Natural Language Inference (SNLI) corpus and use that to generate our sentence representations for both source and target sentences. Authors in [6] show that the sentence encoder achieves the best performance with a Bi-directional LSTM followed by pooling the maximum value over each dimension of the hidden units (max pooling). We choose the training of the sentence encoder on a natural language inference (NLI) dataset because of the strong connection between textual entailment and novelty detection as previously established in the novelty subtask of RTE-TAC [93]. *If a hypothesis is inferred/entailed from a given text, it is usually non-novel. A non-entailed hypothesis, however, may contain new information.* Thus, we deem that the natural language inference task exhibits complex semantic interactions between the source (premise) and target (hypothesis) text pairs required for adjudging the novelty of the target text.

### Encapsulation

We split the documents into constituent sentences. Then encode the sentences into their corresponding embeddings using the SNLI trained BiLSTM+max pooled sentence encoder. For each of the target sentence  $t$  we pull the most similar source sentence  $s$  using cosine similarity. We then encapsulate a target sentence with it's corresponding source as  $ESV_t = [s, t, |s - t|, s * t]$  where  $(,)$  signifies column vector concatenation and  $ESV_t$  is the Encapsulated Sentence Vector of a target sentence  $t$  (Figure 3.8). Finally, for a given target document, we stack the encapsulated sentence representations so obtained to form the Source Encapsulated Target Document Vector (SETDV) matrix. The matrix has a dimension of  $N \times 4D$  where  $N$  is the number of sentences in the target document (padded when necessary), and  $D$  is the sentence embedding dimension produced by the sentence encoder. For this representation we take inspiration from the word embedding studies by [163] where the linear offset of vectors is seen to capture semantic relationships between the two words. Authors in [167] successfully leveraged this idea for modelling sentence-pair relationships which we extend to model source-target relationships in case of documents.

Thus, we arrive at a semantic representation of the target document that has the nearest source information embedded within it. The rationale is that: a novel document would have a different semantic association with the nearest source in the vector space than that of its non-novel counterpart.

### Convolutional Neural Network (CNN)

We use CNN as the feature extractor. Recently CNN has shown great promise in many downstream NLP applications [184]. The document matrix or the SETDV is our input to the CNN

### 3.7 Quantifying Novelty

---

for training and subsequently predicting the novelty score of the target document with respect to the designated set of source documents. We design a CNN similar to [165] used for sentence classification. However, instead of word embeddings as input, we use the source-encapsulated target sentence embeddings of dimension  $4D$  (we represent the  $k^{th}$  sentence in the document by an embedding vector  $ESV_k \in \mathbb{R}^D$ ). We use the NON-STATIC TEXT channel variant of the CNN, where the embeddings get updated during training.

For each possible input channel, a given document is transformed into a tensor of fixed length  $N$  (padded with *zeroes* wherever necessary to tackle variable sentence lengths) by concatenating the relative sentence embeddings.

$$ESV_{1:N} = ESV_1 \oplus ESV_2 \oplus ESV_3 \oplus \dots \oplus ESV_N \quad (3.19)$$

where  $\oplus$  is the concatenation operator. To extract *local features*, a convolution operation is applied. Convolution operation involves a *filter*,  $W \in \mathbb{R}^{HD}$ , which is convolved with a window of  $H$  embeddings to produce a local feature for the  $H$  target sentences. A local feature,  $c_k$  is generated from a window of embeddings  $RSV_{k:k+H-1}$  by applying a non-linear function (*Rectified Linear Unit*) over the convoluted output. Mathematically,

$$c_k = f(W \cdot ESV_{k:k+H-1} + b) \quad (3.20)$$

where  $b \in \mathbb{R}$  is the *bias* and  $f$  is the non-linear function. This operation is applied to each possible window of  $H$  target sentences to produce a feature map ( $c$ ) for the window size  $H$ .

$$c = [c_1, c_2, c_3, \dots, c_{N-H+1}] \quad (3.21)$$

A global feature is then obtained by applying *max-pooling* operation [169] over the feature map. The idea behind *max-pooling* is to capture the most important feature, one with the highest value for each feature map. We describe the process by which we extract one feature from one filter (red filter portion in Figure 3.8 illustrate the case of  $H = 3$ ). The model uses multiple filters for each filter size to obtain multiple features representing the text. These features form the penultimate layer, and we pass them to a fully connected feedforward network (with the number of hidden units set to 100 for the first layer and 50 for the second layer with a dropout of 0.5) followed by a *ReLU* layer whose output is the novelty prediction score.

### 3.7.3 Experiments and Results

We carry on the evaluation on our dataset and automatically predict the novelty score of a document and see how it correlates with human annotated novelty score. Wherever necessary we pad the document representation with zeros.

#### Comparing Systems and Baselines

We design our baselines to serve our ablation study on the proposed model simultaneously. As comparing systems, we cover almost all published works that at any point derives a novelty score.

##### Baseline 1

We leave out SETDV and SNLI pre-training here. We take the *paragraph vector* [161] representation of the target and source documents, concatenate them and pass the joint representation through an MLP. We use the pre-trained *doc2vec* model on a newspaper corpus to generate the embeddings<sup>22</sup>. We select this representation as paragraph vector is known to effectively encode paragraphs/documents leveraging the power of *word2vec* [163].

##### Baseline 2

Next, we went on to investigate the importance of SNLI pre-training and implications of ablating the natural language inference knowledge for novelty detection. Textual Entailment/Natural Language Inference has been known to correlate well with the Novelty Detection [93] task. Hence, instead of taking SNLI trained semantic sentence representations, we generate them using the pre-trained *doc2vec* and use architecture identical to the proposed one (Figure 3.8).

##### Baseline 3

With the third baseline, we want to study how the joint encapsulation of source and target information a.k.a. **SETDV** is crucial to this task. Hence, although we generate sentence representations from pre-trained SNLI, instead of SETDV we stack the sentence representations to form the document representation. We concatenate the three source with the target document representation horizontally and feed them to the subsequent CNN module. Thus except the shape of the input matrix, this baseline resembles the proposed approach.

The intuition behind each of these baselines is to let the network learn the pattern of new and redundant information only from the source and the target data representations.

<sup>22</sup><https://github.com/jhlau/doc2vec>

### 3.7 Quantifying Novelty

---

#### Comparing System 1

As the first comparing system we take the popular *redundancy as opposed to novelty* technique, widely explored in several works [95, 160, 97]. We investigate with both *tf-idf* and *doc2vec* representations of the documents. The reason being although *tf-idf* was the representation used in the original works, we also experiment with the more semantically enriched *doc2vec* to probe the actual effect of the redundancy-distance perspective to novelty scoring. We use the novelty distance metric once in pairwise (*PNov*)

$$PNov(t_i|s_1, \dots s_m) = \min_{1 \leq j \leq m} [1 - \cos(t_i, s_j)] \quad (3.22)$$

and again in aggregate (*ANov*) form [97].

$$ANov(t_i|s_1, \dots s_m) = [1 - \cos(t_i, S_u)] \quad (3.23)$$

where  $S_u = \bigcup_{j=1}^m s_j$ ,  $t_i$  is the target document and  $s_j$  are the source documents.

#### Comparing System 2

We compare with the normalized blended metrics for novelty scoring introduced in [15] using cosine similarity (*cos*) and new word ratio (*nwr*) as the components.

$$J_{blended}(t_i|s_1, \dots s_m) = \alpha J_{nwr}(t_i) + (1 - \alpha) J_{cos}(t_i) \quad (3.24)$$

where  $\alpha$  is the blending parameter ranging from 0 to 1 and is learnt from our training samples ( $\alpha = 0.75$ ).

#### Comparing System 3

We use the minimum Kullback-Leibler (KL) divergence as another comparing system [160, 97]. Thus, the respective novelty scoring formula is as follows:

$$MinKL(t_i|s_1, \dots s_m) = \min_{1 \leq j \leq m} KL(\theta_{t_i}, \theta_{s_j}) \quad (3.25)$$

### Comparing System 4

Authors in [96] proposed a novelty detection algorithm based on *Inverse Document Frequency* scoring function. The novelty score of a new document  $d$  for a collection  $C$  is defined as:

$$NS(d, C) = \frac{1}{\text{norm}(d)} \sum_{q \in d} \text{tf}(q, d) \times \text{idf}(q, C) \quad (3.26)$$

where  $q$  is any term in target document  $d$ ,  $C$  in our case are the designated source documents for  $d$ .

Table 3.9: Performance of the proposed approach against the baselines and comparing systems, PC→ Pearson Correlation Coefficient, MAE→ Mean Absolute Error, RMSE→ Root Mean-Squared Error, Cosine→ Cosine similarity between predicted and actual score vectors

Evaluation System	Description: Novelty Scoring	PC	MAE	RMSE	Cosine
Baseline 1	<i>doc2vec</i> +MLP	0.818	14.027	20.715	0.895
Baseline 2	Without SNLI pre-training	0.834	14.378	19.939	0.902
Baseline 3	Without SETDV encapsulation	0.845	13.686	18.641	0.910
Comparing System 1a	<i>Pairwise: tf-idf</i> [185, 160]	0.029	32.441	37.161	0.734
Comparing System 1b	<i>Pairwise: doc2vec</i>	0.347	40.993	54.315	0.782
Comparing System 1c	<i>Aggregate: tf-idf</i> [97]	0.130	32.281	38.901	0.728
Comparing System 1d	<i>Aggregate: doc2vec</i>	0.494	41.004	54.347	0.809
Comparing System 2a	<i>Blended</i> [92]	0.680	23.733	28.202	0.870
Comparing System 2b	<i>Blended using doc2vec</i>	0.685	40.990	54.351	0.871
Comparing System 3	Min. KLD [160]	0.592	35.997	47.718	0.846
Comparing System 4	Inverse Document Frequency [96]	0.160	41.236	54.671	0.576
<b>Proposed Approach</b>	<b>SETDV-CNN</b>	<b>0.888</b>	<b>10.294</b>	<b>16.547</b>	<b>0.953</b>

### 3.7.4 Results and Discussion

We deduce the novelty score of each target document with our SETDV-CNN architecture. Performance comparison of our approach with the baselines and *state-of-the-arts* are presented in Table 3.9. It is quite evident that SETDV-CNN is performing way better than the baselines and the comparing systems. The reason for the low performance of the comparing systems is because those were mostly designed from an IR perspective and did not address the semantic-level information needs. However, baselines came close to the proposed approach as they manifest enriched semantic vector composition from which we extract features via neural networks.

**Baseline 1** performs comparatively poor as we ablate both SNLI pre-training of the sentence vectors as well as the SETDV-CNN. In **Baseline 2** when we ablate the SNLI pre-training but keep the SETDV-CNN framework, we gain a little improvement. **Baseline 3** came out as the strongest with only the SNLI pre-training preserved. This indicates that the inference knowledge gained from training on SNLI is an important component to understand the notion of novelty. However, the higher performance of the proposed method and the adopted baselines



### 3.7 Quantifying Novelty

---

for document-level novelty scoring clearly indicate that deep neural networks are efficient than existing feature-based and rule-based techniques for the problem under study.

We also experiment with a variant of Comparing Systems 1(b,d) and 2(b) using the semantically enriched *doc2vec* representation which supposedly gives better performance than *tf-idf* (See in Table 3.9). It's interesting to see that our stronger baselines perform better than the *state-of-the-art's* which seconds the proposition that incorporating semantic knowledge actually improves the prediction performance. We also find that our method is more prone towards discovering redundant information, i.e. documents having low novelty score This is good considering that the dataset exhibits semantic-level redundancy. Usually, novel texts differ in lexical-level as well and hence are easier to identify. The actual challenge lies in identifying the semantically redundant textual content where the existing methods score low. Our method is efficient as we observe a higher correlation between the actual scores (human-annotated gold standard) and our system predicted scores in the scatter plot in Figure 3.9.

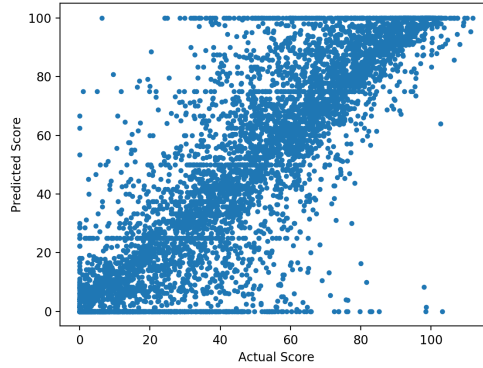


Figure 3.9: Scatter plot of Actual (Gold Standard) vs Predicted (Proposed) Document-Level Novelty Score

#### 3.7.5 Error Analysis

We analyse the predictions and identify the following class of errors committed by our system:

- When the target document is comparatively too small with respect to the source and other documents in the dataset. A significant amount of padding with zeros results in affinity towards non-novelty.
- **Multiple premise scenarios:** This is when a target sentence derives information from multiple source sentences. Hence, selecting only a single sentence as the source goes against

our annotation perspective (during annotation we consider the overall knowledge gained from reading the source documents, not a specific source text).

- Target document having a complex syntactic structure as compared to the source and difficult to comprehend as well (e.g., too many complex and compound sentences).
- Target document has a different narrative style w.r.t. source (e.g., active vs passive voice). Such syntactic nuances were not captured correctly by our sentence encodings.
- Annotation conflicts among the annotators caused some errors. This happens mostly because of the (i) different novelty appetite of the annotators and (ii) not considering role of sentence significance within a document discourse for judging new information.
- Persistent noises, error in sentence splitting prohibited to form a complete semantic unit (Figure 3.2).

Here in this work we discuss our approach towards quantifying the *newness* in a document. In our next exploration we investigate *multi-premise inference* for the problem.

## 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

Non-novel/redundant information in a document may have assimilated from multiple source documents, not just one. The problem surmounts when the subject of the discourse is documents, and numerous prior documents need to be processed to ascertain the novelty/non-novelty of the current one in concern. In this work, we build upon our earlier investigations for document-level novelty detection and present a comprehensive account of our efforts towards the problem. We explore the role of pre-trained Textual Entailment (TE) models to deal with multiple source contexts and present the outcome of our current investigations. We argue that a multi-premise entailment task is one close approximation towards identifying semantic-level non-novelty. Our recent approach either performs comparably or achieves significant improvement over the latest reported results on several datasets and across several related tasks (paraphrasing, plagiarism, rewrite). We argue that to ascertain a given text’s novelty, we would need multi-hop reasoning on the source texts for which we draw reference from the question-answering literature [186].

### 3.8.1 Textual Novelty Detection: An Entailment Perspective

Textual Entailment is defined as a directional relationship between two text fragments, termed Text (T) and Hypothesis (H) as:

T ENTAILS H IF, TYPICALLY, A HUMAN READING T WOULD INFER THAT H IS MOST LIKELY TRUE [187].

For example, let us consider the following two texts:

#### Example 1

**Text 1:** *I left the restaurant satisfactorily.* (Premise **P**)

**Text 2:** *I had good food.* (Hypothesis **H**)

So a human reading Text 1 (Premise) would most likely infer that Text 2 (Hypothesis) is true, i.e., Text 1 entails Text 2 or the Premise P entails the Hypothesis H.

The PASCAL-RTE challenges [188, 189] associated textual novelty with entailment. As RTE puts: *RTE systems are required to judge whether the information contained in each  $H$  is novel with respect to (i.e., not entailed by) the information contained in the corpus. If entailing sentences ( $T$ ) are found for a given  $H$ , it means that the content of the  $H$  is not new (**redundant**); in contrast, if no entailing sentences are detected, it means that information contained in  $H$  is **novel**.* With respect to the above example, we can say that Text 1 is known to us in a specific context, Text 2 probably has no new information to offer. However, there could be other reasons for one leaving the restaurant satisfactorily:

- The ambience was good ( $H_1$ )
- The price was low ( $H_2$ )
- I got some extra fries at no cost ( $H_3$ )
- I received my birthday discount at the restaurant ( $H_4$ )

etc. However, the probability of inferring  $H_1, H_2, H_3, H_4$  given  $P$  seems relatively low as compared to inferring  $H$  given  $P$  in a general context.

$$Pr(H|P) > Pr(H_1|P, H_2|P, H_3|P, H_4|P)$$

Rather, we say that given  $P$ , we can implicitly assume with a higher degree of confidence that  $H$  is true. So,  $H$  might not be offering any new information. However, the same cannot be postulated for  $H_1, H_2, H_3, H_4$  given  $P$ . Hence, the probability of  $H$  being *non-novel* given  $P$  is higher than  $H_1$  given  $P, H_2$  given  $P, H_3$  given  $P, H_4$  given  $P$ . Having said that, without a given context,  $H_1, H_2, H_3, H_4$  are probably offering some relatively *new* information with respect to the premise  $P$ . Please note that there is a minimum lexical overlap between the Premise and the Hypothesis texts. The overlap is at the semantic level. Supposedly, textual entailment at the semantic level is more close to the detection of non-novelty.

Inspired by this idea of associating entailment probabilities to texts with respect to premises, we went on to explore how we could train a machine learning architecture to identify the novelty of not only a single sentence but for an entire document. However, our investigation is different from earlier explorations in the sense that:

- Novelty detection tasks in both the TREC [20] and RTE-TAC [189] were designed from an information retrieval perspective where the main goal was to retrieve relevant sentences

### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

---

to decide on the novelty of a statement. We focus on automatic classification/scoring of a document based on its new information content from a machine learning perspective.

- As is evident from the examples, the premise-hypothesis pair shows significantly less lexical overlap, making the entailment decisions more challenging while working at the semantic level. Our methods encompass such semantic phenomena, which was less prominent in the TREC and RTE-TAC datasets.
- For ascertaining the novelty of a statement, we opine that a single premise is not enough. We would need the context, world knowledge, and reasoning over multiple facts. We discuss the same in the subsequent section.

#### 3.8.2 Multi-premise Entailment for Novelty Detection

We deem the NLP task multi-premise entailment as one close approximation to simulate the phenomenon of textual non-novelty. Multi-premise Entailment (MPE) [171] is a variant of the standard Textual Entailment task in which the premise text consists of multiple independently written sentences (source), all related to the same topic. The task is to decide whether the hypothesis sentence (target) can be used to describe the same topic (entailment) or cannot be used to describe the same topic (contradiction), or may or may not describe the same topic (neutral). The main challenge is to infer what happened in the topic from the multiple premise statements, in some cases aggregating information across multiple sentences into a coherent whole. The MPE task is more pragmatic than the usual Textual Entailment as it aims to assimilate information from multiple sources to decide the entailment status of the hypothesis.

Similarly, the novelty detection problem becomes more practical and hence intense when we need to consider multiple sources of knowledge (premises) to arrive at a decision whether a given text (hypothesis) contains new information or not. In the real world, it is highly unlikely that a certain text would assimilate information from just another text (unlike the Premise-Hypothesis pair instances in most Natural Language Inference (NLI) datasets). For deciding on the novelty of a text, we would need to consider the context and reason over multiple facts. Let us consider the following example. Here, *source* would signify information that is already seen or known (Premise) to the reader, and *target* would signify the text for which novelty/redundancy is to be ascertained (Hypothesis).

#### Example 2

**Source:** *Survey says Facebook is still the most popular social networking site ( $s_1$ ). It was created by Mark Zuckerberg and his colleagues when they were students at Harvard back in 2004 ( $s_2$ ). Harvard University is located in Massachusetts, Cambridge, which is just a few miles from Boston ( $s_3$ ). Zuckerberg now lives in Palo Alto, California ( $s_4$ ).*

**Target:** *Facebook was launched in Cambridge ( $t_1$ ). The founder resides in California ( $t_2$ ).*

Clearly, the target text would appear *non-novel* to a reader with respect to the premise. However, to decide on each sentence’s novelty in the target text, we would need to consider multiple sentences in the source text, not just one. Here in this case, to decide on the novelty of  $t_1$ , we would need the premises  $s_1$ ,  $s_2$ ,  $s_3$  and similarly  $s_1$ ,  $s_2$ ,  $s_4$  to decide for  $t_2$ .  $s_4$  is not of interest for  $t_1$ , neither is  $s_3$  for  $t_2$ . Thus to answer for the novelty of a certain text, it is quite likely that we may need to reason over multiple relevant sentences. Hence a multi-premise inference scenario appears to suit here. In our earlier work [2], we already consider *Relevance* to be one important criteria for *Novelty Detection*. So, selecting relevant premises for a statement is an important step towards detecting the novelty of the statement.

### 3.8.3 Encompassing Multiple Premises for Document-Level Novelty Detection

As we discuss earlier, reasoning over multiple facts is essential for textual novelty detection. We may need to assimilate information from multiple source texts to ascertain the state of the novelty of a given statement or a fact. *If a text is redundant against a given prior, it is redundant against the set of all the relevant priors. However, it has to be novel against all the relevant priors for a text to be novel.* Here, a prior signifies the relevant information exposed to the reader that s/he should refer to determine the *newness* of the *target text*. If no such priors are available, possibly the target text has new information. Organizers of TREC information retrieval exercises [83] formulated the tasks along this line. If for a given query (target), no relevant source is found from a test collection, possibly the query is new. Here  $s_1$ ,  $s_2$ ,  $s_3$ ,  $s_4$  are the relevant priors for  $t_1$ ,  $t_2$ .

We also indicate in our earlier work [65] that the selection of relevant prior information is an important precursor towards deciding the novelty of a given statement or fact. Hence, finding the relevant source sentences is essential towards ascertaining the *newness* of the target sentence. Hence, in our proposed approach, we encompass two components:

- a relevance detection module, followed by

### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

- a novelty detection module

We make use of pre-trained natural language inference models (entailment models) for both the components. To assimilate information from multiple priors, the novelty detection module manifests a *join* operation at multiple layers of the pre-trained entailment stack to capture multiple levels of abstraction. The join operation is inspired from [190] for question-answering. It results in a multi-premise (source) aware hypothesis (target) representation, where we combine all such target sentence representations to decide on the novelty of the target document. Figure 3.10 shows the architecture of our proposed approach.

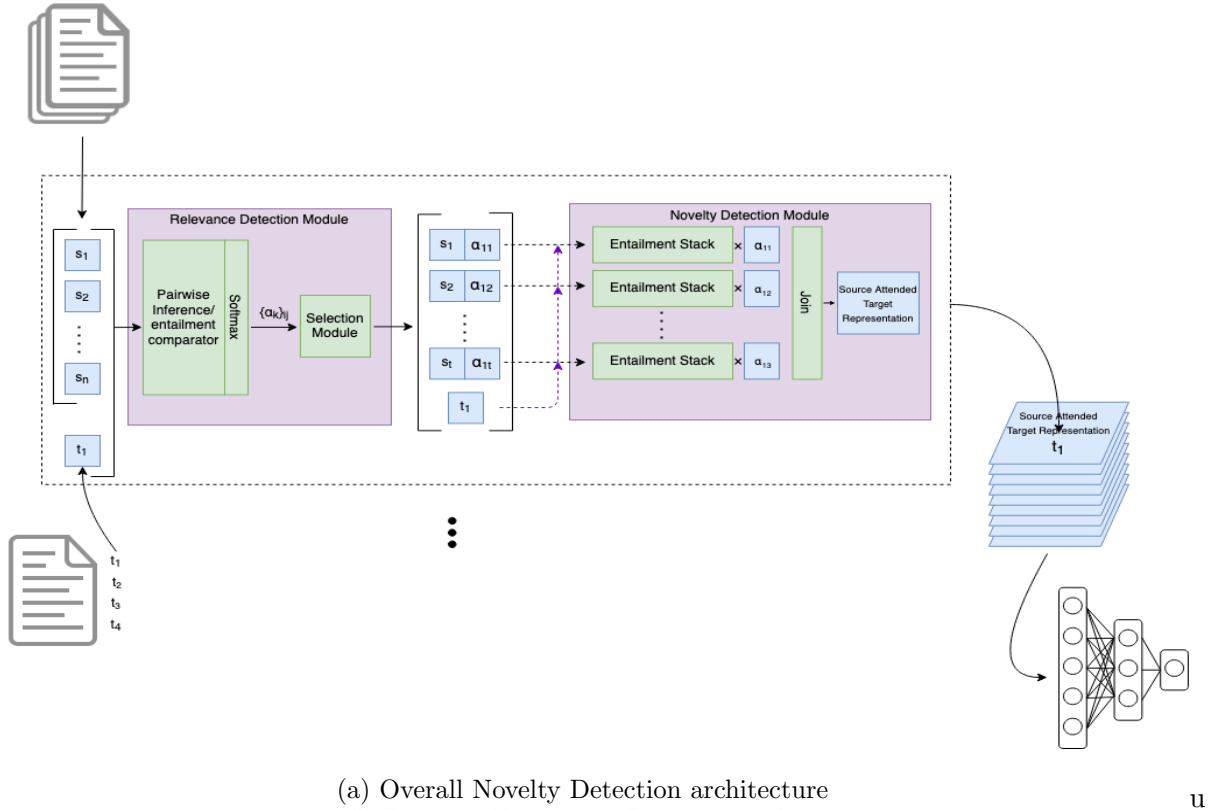


Figure 3.10: Multi-Premise Entailment Based Document-Level Novelty Detection Architecture

## Relevance Detection

The goal of this module is to find relevant premises (source sentences) for each sentence in the target document. We treat the sentences in the target document as our multiple hypotheses, i.e., we see a target document to comprise multiple hypothesis statements. The objective is to find to what extent each of these hypotheses is entailed from the premises in the source documents and use that knowledge to decide the target document’s novelty. Ideally, *a non-novel document would find majority of its sentences highly entailed from the various sentences in the source documents*. A source sentence is considered relevant if it contains information related to the target sentence and may serve as the premise to determine the newness of the target sentence. We model this relevance in terms of entailment probabilities, i.e., how well the information in the source and the target correlates. We use a pre-trained inference model to give us the entailment probabilities between all possible pairs of target and source sentences. Not all sentences in the source documents would be relevant for a given target sentence (as per the example in the earlier section,  $s_4$  is not relevant for  $t_1$  and  $s_3$  is not relevant to  $t_2$ ). For each target sentence ( $t_k$ ), we select the top  $f$  source sentences with the highest entailment probabilities ( $\alpha_{kf}$ ) as the relevant priors. After softmax, the final layer of a pre-trained entailment model would give us the entailment probability between a given premise-hypothesis pair.

## Input

Let  $S_1, S_2, \dots, S_n$  are the source documents retrieved from a document collection for a target document  $T$ . In our experiments we already had the source documents designated for a given target document. We split the source and target documents into corresponding sentences. Here,  $s_{ij}$  denotes the  $i$ th sentence of the source document  $j$ .  $t_k$  represents the sentences in the target document ( $T$ ). The final objective is to determine whether  $T$  is *novel* or *non-novel* with respect to  $S_1, S_2, \dots, S_n$ .

## Inference Model

The source-target sentence pairs are then fed to a pre-trained Natural Language Inference (NLI) model to get the entailment probabilities after the softmax layer. Here, we make use of the Enhanced Sequential Inference Model (ESIM) [191] trained on large-scale inference datasets: SNLI [192] and MultiNLI [193] as our pre-trained entailment stack.

$$\{\alpha_k\}_{ij} := Pr[s_{ij} \rightarrow t_k] \tag{3.27}$$



### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

---

where  $\{\alpha_k\}_{ij}$  denotes probability of entailing  $t_k$  from source sentence  $s_{ij}$ . This is the output of the pre-trained ESIM model’s softmax layer on Premise  $s_{ij}$  and Hypothesis  $t_k$ .

#### Selection Module and Relevance Scores

Not all the source sentence would contribute towards the target sentence. Hence we retain the topmost  $f$  relevant source sentences for the target sentence  $t_k$  based on the entailment probabilities or what we term as the *relevance scores*. In Figure 3.10,  $\alpha_{kf}$  denotes the relevance scores for the top  $f$  selected source sentences for a target sentence  $t_k$ . We would further use these relevance scores while arriving at a Source-Aware Target (SAT) representation in the Novelty Detection module. Thus, the relevance module’s outputs are multiple relevant source sentences  $s_{kf}$  for a given target sentence  $t_k$  and their pairwise relevance scores.

#### Novelty Detection Module

The goal of the Novelty Detection module is to assimilate information from the multiple relevant source sentences (from source documents) to ascertain the novelty of the target document. The novelty detection module would take as input the target document sentences paired with their corresponding  $f$  relevant source sentences. This module would again make use of a pre-trained entailment model (ESIM here) along with the *relevance scores* between each source-target sentence pairs from the earlier module to independently arrive at a Source-Aware Target (SAT) representation for each target sentence  $t_k$ . We use the earlier module’s relevance scores to incentivize the contributing source sentences and penalize the less-relevant ones for the concerned target sentence. Finally, we concatenate the  $k$  Source-Aware Target (SAT) representations, pass through a final feedforward and linear layer, to decide on the novelty of  $T$ . We discuss the assimilation of multiple premises weighted by their relevance scores in the following section. The number of entailment functions in this layer depends on the number of target sentences ( $k$ ) and the number of relevant source sentences you want to retain for each target sentence (i.e.,  $f$ ).

#### Relevance-weighted inference model to support multi-premise entailment

A typical neural entailment model consists of an input encoding layer, local inference layer, and inference composition layer (see Figure 3.10b). The input layer encodes the premise (source) and hypothesis (target) texts; the local inference layer makes use of cross-attention between the premise and hypothesis representations to yield entailment relations, followed by additional layers that use this cross-attention to generate premise attended representations of the hypothesis and vice versa. The final layers are classification layers, which determine entailment based on the representations from the previous layer. In order to assimilate information from multiple source

sentences, we use the *relevance scores* from the previous module to scale up the representations from the various layers of the pre-trained entailment model (E) and apply a suitable *join* operation [190]. In this join operation, we use a part of the entailment stack to give us a representation for each sentence pair that represents important features of the sentence pair and hence gives us a meaningful document level representation when combined with weights. We denote this part of the stack as  $f_{e1}$ . The rest of the entailment stack that we left out in the previous step is used to obtain the final representation from the combined intermediate representations and is denoted by  $f_{e2}$ . This way, we aim to emphasize the top relevant source-target pairs and attach lesser relevance scores to the bottom ones for a given target sentence  $t_k$ . The *join* operation would facilitate the assimilation of multiple source information to infer on the target.

We now discuss how we incorporate the *relevance scores* to various layers of the pre-trained entailment model (E) and assimilate the multiple source information for a given target sentence  $t_k$ .

### Input Layer to Entailment Model

For convenience, let us denote any source sentence (premise) as  $\mathbf{s}$  and any target sentence (hypothesis) as  $\mathbf{t}$ .

$$\begin{aligned} s &= (x_1, x_2, x_3, \dots, x_{l_s}) \\ t &= (y_1, y_2, y_3, \dots, y_{l_t}) \end{aligned} \tag{3.28}$$

where  $x_1, x_2, x_3, \dots$  are tokens of source sentence  $\mathbf{s}$  and  $y_1, y_2, y_3, \dots$  are tokens of target sentence  $\mathbf{t}$ . The length of  $\mathbf{s}$  and  $\mathbf{t}$  are  $l_s$  and  $l_t$  respectively.

There is a BiLSTM encoder to get the representation of  $\mathbf{s}$  and  $\mathbf{t}$  as:

$$\begin{aligned} \bar{s}_i &= \{BiLSTM(s)\}_i, i \in (1, 2, \dots, l_s) \\ \bar{t}_j &= \{BiLSTM(t)\}_j, j \in (1, 2, \dots, l_t) \end{aligned} \tag{3.29}$$

where  $\bar{s}_i$  denotes the output vector of BiLSTM at the position  $i$  of the premise, which encodes word  $s_i$  and its context.

### Cross-Attention Layer

Next, is the cross attention between the source and target sentences to yield the entailment relationships. In order to put emphasis on the most relevant source-target pairs, we scale the cross-

### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

---

attention matrices with the *relevance scores* from the previous module and then re-normalize the final matrix.

Cross-attention between source to target and target to source is defined as:

$$\begin{aligned}\tilde{s}_i &= \sum_{j=1}^{l_t} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_t} \exp(e_{ik})} \bar{t}_j \\ \tilde{t}_j &= \sum_{i=1}^{l_s} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_s} \exp(e_{jk})} \bar{a}_i\end{aligned}\tag{3.30}$$

where,  $e_{ij} = (\bar{s}_i)^T \bar{t}_j$

So, for a source sentence  $s$  against a given target sentence  $t$ , we get a source to target cross-attention matrix  $\tilde{s}_i$  and a target to source cross-attention matrix  $\tilde{t}_j$  with dimension  $(i \times j)$  and  $(j \times i)$  respectively.

Now for our current multi-source and multi-target scenario, let for the given target sentence  $t_k$ , we found  $f$  relevant source sentences  $s_{k1}, s_{k2}, \dots, s_{kf}$ . The assimilation mechanism would scale the corresponding attention matrices by a factor  $\alpha_{kf}$  for each source ( $s_f$ )-target ( $t_k$ ) pair to generate the Source-Attended Target Representation (SAT) for  $t_k$  against  $s_{k1}, s_{k2}, \dots, s_{kf}$ .

We scale the cross-attention matrix with the *relevance scores* ( $\alpha_{kf}$ ) to prioritize the important source sentences for a given target sentence and concatenate the matrices for all the  $f$  source sentences ( $s_{k1}, s_{k2}, \dots, s_{kf}$ ) against a given target sentence  $t_k$ .

$$\tilde{s}_i^{s_{kf}t_k} = [\alpha_{k1}\tilde{s}_i^{s_{k1}t_k}; \dots; \alpha_{kf}\tilde{s}_i^{s_{kf}t_k}]\tag{3.31}$$

where  $k$  remains unchanged for a given  $t_k$  and  $f$  varies for the multiple source sentences against a given  $t_k$ .

We concatenate the source sentences ( $s_{k1}, s_{k2}, \dots, s_{kf}$ ) for a given  $t_k$  to obtain the passage-level representation as:

$$[S_{kf}] = [[\alpha_{k1}\bar{s}_{k1}]; [\alpha_{k2}\bar{s}_{k2}]; \dots; [\alpha_{kf}\bar{s}_{kf}]]\tag{3.32}$$

We keep the target sentence representation ( $\bar{t}_k$ ) unchanged. We forward the scaled attention matrices, scaled source representations, and the unchanged target representation to the next layer in the entailment stack. We repeat the same operation for all the sentences ( $t_1, t_2, \dots, t_k$ ) in the target document  $T$ .

### Source-Aware Target Representations

We also scale the final layer in the entailment stack ( $E_{kf}$ ) with the *relevance scores* ( $\alpha_{kf}$ ). The final layer in the entailment stack usually outputs a single vector  $\bar{h}$ , which is then used in a linear layer and a final logit to obtain the final decision. The join operation here is a weighted sum of the source-target representations from the preceding layers. So we have:

$$SAT_k = \sum_f \alpha_{kf} h_{kf} \quad (3.33)$$

where  $SAT_k$  is the Source-Aware Target representation for  $t_k$ . We do the same for all the target sentences in the target document  $T$ .

### Novelty Classification

We stack the Source-Aware Target representations ( $SAT_k$ ) for all the sentences in the target document and pass the fused representation through a Multi-Layer Perceptron (MLP) to discover important features and finally classify with a softmax layer. The output is whether the target document is *Novel* or *Non-Novel* with respect to the source documents.

#### 3.8.4 Datasets for Allied Tasks

Finding semantic-level redundancy is challenging than finding novelty in texts [64]. The challenge scales up when it is at the level of documents. Semantic-level redundancy is a good approximation of non-novelty. Novel texts usually consist of new terms and generally are lexically different from the source texts. Hence with our experiments, we stress on detecting non-novelities, which would eventually lead us to identify novelties in text. Certain tasks could simulate the detection of non-novelty. Paraphrasing is one such linguistic task where paraphrases convey the same information as the source texts yet have a very less lexical similarity. Another task that comes close to identifying novelties in the text is Plagiarism detection, which is a common problem in academia. We train our model with the document-level novelty datasets and test its efficacy to detect paraphrases and plagiarized texts. We employ the following well-known datasets for our investigation.

#### Webis Crowd Paraphrase Corpus

The Webis Crowd Paraphrase Corpus 2011 (Webis-CPC-11) [194] consists of 7,859 candidate paraphrases obtained from the Mechanical Turk crowdsourcing. The corpus<sup>23</sup> is made up of

---

<sup>23</sup><https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/data/corpus>

### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

4,067 accepted paraphrases, 3,792 rejected non-paraphrases, and the original texts. For our experiment, we assume the original text as the source document and the corresponding candidate paraphrase/non-paraphrase as the target document. We hypothesize that a paraphrased document would not contain any new information, and we treat them as *non-novel* instances. Table 3.10 shows an example of our interpretation of non-novelty in the dataset.

Table 3.10: Sample text from Webis-CPC-11 to simulate the high-level semantic paraphrasing in the dataset.

Original Text (Source Document)	Paraphrase Text (Target Document: Non-Novel)
The emigrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships. Accordingly, the Swallow was sent back to England with the sick; and with the remainder of the fleet, well supplied at St. John's with fish and other necessities, Gilbert (August 20) sailed south as far as forty-four degrees north latitude. Off Sable Island a storm assailed them, and the largest of the vessels, called the Delight, carrying most of the provisions, was driven on a rock and went to pieces.	The people who left their countries and sailed with Gilbert were more suited for fighting the crusades than for leading a settled life in the colonies. They were bitterly disappointed as it was not the America that they had expected. Since they did not immediately find gold and silver mines, many deserted. At one stage, there were not even enough men to help sail the four ships. So the Swallow was sent back to England carrying the sick. The other fleet was supplied with fish and the other necessities from St. John. On August 20, Gilbert had sailed as far as forty-four degrees to the north latitude. His ship known as the Delight, which bore all the required supplies, was attacked by a violent storm near Sable Island. The storm had driven it into a rock shattering it into pieces.

### P4PIN Plagiarism Corpus

We use the P4PIN corpus [195], a corpus especially built for evaluating the identification of paraphrase plagiarism. This corpus is an extension of the P4P corpus [196], which contains pairs of text fragments where one fragment represents the original source text, and the other represents a paraphrased version of the original. In addition, the P4PIN corpus includes *not paraphrase* plagiarism cases, i.e., negative examples formed by pairs of unrelated texts samples with likely thematic or stylistic similarity. The P4PIN dataset consists of 3,354 instances, 847 positives, and 2,507 negatives. We are interested in detecting plagiarism cases and also see the novelty-scores for each category of instances predicted by our model. Table 3.11 shows a plagiarism (non-novel) example from P4PIN.

### Wikipedia Rewrite Corpus

The dataset [197] contains 100 pairs of short texts (193 words on average). For each of 5 questions about topics of computer science (e.g., "What is dynamic programming?"), a reference answer (source text, hereafter) has been manually created by copying portions of text from a relevant

Table 3.11: Sample from P4PIN to show plagiarism (non-novel) instance

Original Text (Source Document)	Plagiarised Text (Target Document: Non-Novel)
I pored through these pages, and as I perused the lyrics of The Unknown Eros that I had never read before, I appeared to have found out something wonderful: there before me was an entire shining and calming extract of verses that were like a new universe to me.	I dipped into these pages, and as I read for the first time some of the odes of The Unknown Eros, I seemed to have made a great discovery: here was a whole glittering and peaceful tract of poetry, which was like a new world to me.

Wikipedia article. According to the degree of the rewrite, the dataset is 4-way classified as *cut & paste* (38 texts; a simple copy of text portions from the Wikipedia article), *light revision* (19; synonym substitutions and changes of grammatical structure allowed), *heavy revision* (19; rephrasing of Wikipedia excerpts using different words and structure), and *no plagiarism* (19; answer written independently from the Wikipedia article). We test our model on this corpus to examine the novelty-scores predicted by our proposed approach for each category of answers. Please note that the information content for each of these answer categories are more or less the same as they cater to the same question.

Table 3.12: Sample from Wikipedia Rewrite Dataset to show a plagiarism (non-novel) instance

Original Text (Source Document)	Plagiarised Text (Target Document: Non-Novel)
PageRank is a link analysis algorithm used by the Google Internet search engine that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of “measuring” its relative importance within the set.	The PageRank algorithm is used to designate every aspect of a set of hyperlinked documents with a numerical weighting. It is used by the Google search engine to estimate the relative importance of a web page according to this weighting.

### 3.8.5 Evaluation

In this section we evaluate the performance of our proposed approach, compare with baselines and also with our earlier approaches. We further show how our model performs in allied tasks like paraphrase detection, plagiarism detection, and identifying rewrites.

#### Baselines

We carefully choose our baselines so that those also helps in our ablation study.

#### Baseline 1: Joint Encoding of Source and Target Documents

With this baseline, we want to see the importance of textual entailment for our task of textual novelty detection. We use the Transformer variant of the Universal Sentence Encoder [198] to encode sentences in the documents to fixed-sized sentence embeddings (512 dimension) and

### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

---

then stack them up to form the document embedding. We pass the source and target document representations to a Multi-layer Perceptron for corresponding feature extraction and final classification via *Softmax*.

#### Baseline 2: Importance of Relevance Detection

With this baseline, we investigate the significance of relevance detection as a prior task to novelty detection. We turn off the relevance detection module and use the individual entailment decisions from the pre-trained ESIM model to arrive at the document-level aggregated decision.

#### Baseline 3: Single Premise

We keep all other parameters of our proposed model intact, but instead of having multiple premise, we take only the closest (top) premise (from the source sentences) for each target sentence. This way we want to establish the importance of aggregating multi-premise entailment decisions for document-level novelty detection.

#### Comparing Systems

We compare with our earlier works on the same datasets, keeping all experimental configurations the same. Kindly refer to our earlier papers for detailed overview of the techniques.

#### Comparing System 1

With our first exploration on document-level novelty detection [2] we use several features ranging from *lexical similarity*, *semantic similarity*, *divergence*, *keywords/named-entities overlap*, *new word count*, *etc..* The best-performing classifier was Random Forest [199].

#### Comparing System 2

With our next exploration, we introduce the concept of a Relative Document Vector (RDV) as a fused representation of source and target documents [64]. We use a Convolutional Neural Network (CNN) to extract useful features for classifying the target document into novelty classes.

#### Comparing System 3

To determine the amount of new information (novelty score) in a document, we generate a Source-Encapsulated Target Document Vector (SETDV) and train a CNN to predict the novelty score of the document [65]. The value of the novelty score of a document ranges between 0 to 100 on the basis of new information content as annotated by our annotators.

## Comparing System 4

With this work we went on to explore the role of textual alignment (via decomposable attention mechanism) between target and source documents to produce a joint representation [200]. We use a feed forward network to extract features and classify the target document on the basis of new information content.

## Hyperparameter Details

Our current architecture uses ESIM stack as the entailment model pre-trained on SNLI and MultiNLI for both the relevance module and for the novelty detection module. Binary Cross Entropy is the loss function and the default dropout is 0.5. We train for 10 epochs with *Adam* optimizer and keep the learning rate as 0.0004. The final feedforward network has *ReLU* activations with a dropout of 0.2. The input size for the Bi-LSTM context encoder is 300 dimension. We use the GloVe 800B embeddings for the input tokens. For all uses of ESIM in our architecture we initialize with the same pre-trained entailment model weights available with AllenNLP [201].

### 3.8.6 Results

We discuss the results of our current approach in this section. We use TAP-DLND 1.0 and APWSJ datasets for our novelty classification experiments and the proposed TAP-DLND 1.1 dataset for quantifying new information experiments. We also report our experimental results on Webis-CPC dataset where we assume paraphrases to be simulating semantic-level non-novelty. We also show use cases of our approach for semantic-level plagiarism detection (other form of non-novelty in academia) with P4PIN and Wikipedia Rewrite datasets.

## Evaluation Metrics

For the novelty classification task, we keep the usual classification metrics: Precision, Recall,  $F_1$  score, and Accuracy. For the APWSJ dataset, instead of accuracy we report the Mistake (100-Accuracy) to do comparison with the earlier works. For the novelty scoring experiments on TAP-DLND 1.1 we evaluate our baselines and proposed model against the ground-truth scores using Pearson Correlation Co-efficient, Mean Absolute Error (the lower the better), Root Mean Squared Error (the lower the better), and the Cosine similarity between the actual scores and the predicted scores.



### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

---

#### On TAP-DLND 1.0 Dataset

Table 3.13 shows our results on TAP-DLND 1.0 dataset for the novelty classification task. As discussed earlier, here we keep  $f = 10$ , i.e. topmost ten relevant source sentences (based on  $\alpha_{kf}$  scores) as the relevant premises for each target sentence  $t_k$  in the target document. We could see that our current approach performs comparably with our preceding approach (Comparing System 4). With a high recall for non-novel class we can say that our approach has an affinity to discover document-level *non-novelty* which is comparatively more challenging at the semantic-level. The results in Table 3.13 are from 10-*fold* cross-validation experiments.

Table 3.13: Results on TAP-DLND 1.0 , P→ Precision, R→ Recall, A→ Accuracy, R→ Recall, N→ Novel, NN→ Non-Novel, 10-*fold* cross-validation output

Evaluation System	P(N)	R(N)	$F_1$ (N)	P(NN)	R(NN)	$F_1$ (NN)	A
Baseline 1	0.61	0.77	0.67	0.53	0.57	0.55	68.1%
Baseline 2	0.84	0.57	0.67	0.71	0.86	0.77	76.4%
Baseline 3	0.82	0.70	0.76	0.77	0.84	0.80	80.3%
Comparing System 1	0.77	0.82	0.79	0.80	0.76	0.78	79.3%
Comparing System 2	0.86	0.87	0.86	0.84	0.83	0.83	84.5%
Comparing System 4	0.85	0.85	0.85	0.89	0.89	0.89	87.4%
<b>Proposed Approach</b>	0.94	0.77	0.85	0.80	0.95	0.87	87.2%

#### On APWSJ Dataset

The APWSJ dataset is more challenging than TAP-DLND 1.0 because of the sheer number of preceding documents one has to process for deciding the state of novelty of the current one. The first document in the chronologically ordered set of documents for a given topic is always *novel* as it starts the story. The novelty of all other documents are judged based on the chronologically preceding ones. So for the final document in a given topic, the network needs to process all the preceding documents in that topic. Although, APWSJ was developed from an information retrieval perspective, we take a classification perspective (i.e. to classify the current document into *novel* or *no-novel categories* based on its chronological priors) for our experiments. Table 3.14 reports our result and compares with earlier systems. Kindly note that we take the same experimental condition as the original paper [3] and consider *partially-redundant* documents into the *redundant* class. Our current approach perform much better than the earlier reported results with  $f = 10$ , thereby signifying the importance of multi-premise entailment for the task in hand. We report our results on the redundant class as in earlier systems. Finding semantic-level non-novelty for documents is much more challenging than identifying whether a document has enough new things to say to classify it as *novel*.

Table 3.14: Results for redundant class on APWSJ, *Mistake*  $\rightarrow$  100-Accuracy. Except for [3] all other figures correspond to a 10-*fold* cross-validation output

Measure	Recall	Precision	Mistake
Baseline 1	0.66	0.75	28.8%
Baseline 2	0.76	0.85	18.8%
Baseline 3	0.85	0.86	13.4%
Comparing System [3]	0.56	0.67	27.4%
Comparing System 2	0.58	0.76	22.9%
Comparing System 4	0.86	0.92	7.8%
<b>Proposed Approach</b>	<b>0.91</b>	<b>0.95</b>	<b>5.9%</b>

### On TAP-DLND 2.0 Dataset

On our newly created dataset for predicting novelty-scores, instead of classification we try to squash the output to a numerical score. We use the same architecture in Figure 3.10 but use *sigmoid* activation at the last layer to restrict the score within the range of 100. Table 3.15 shows our performance. This experiment is particularly important to quantify the amount of *newness* in the target document with respect to the source documents. Kindly note we allow a +5 and -5 range with respect to the human-annotated score for our predicted scores. We see that our current approach performs comparably with the earlier reported results.

Table 3.15: Performance of the proposed approach against the baselines and comparing systems TAP-DLND 1.1, PC $\rightarrow$  Pearson Correlation Coefficient, MAE $\rightarrow$  Mean Absolute Error, RMSE $\rightarrow$  Root Mean-Squared Error, Cosine $\rightarrow$  Cosine similarity between predicted and actual score vectors

Evaluation System	PC	MAE	RMSE	Cosine
Baseline 1	0.69	36.11	49.92	0.87
Baseline 2	0.81	15.34	23.83	0.91
Baseline 3	0.84	12.40	20.14	0.93
Comparing System 3	0.88	10.29	16.54	0.95
<b>Proposed Approach</b>	<b>0.88</b>	<b>10.92</b>	<b>17.73</b>	<b>0.94</b>

### Ablation Studies

As we mention, our baselines serves as means of ablation studies. Baseline 1 is the simplest one where we simply let the network to discover useful features from the universal representations of the *source-target* pairs. We do not employ any sophisticated approach, and it performs the worse. Baseline 1 establishes the importance of our textual-entailment pipeline in the task. In Baseline 2, we do not consider the *relevance* detection module and hence do not include the relevance weights in the architecture. Baseline 2 performs much better than Baseline 1 (relative improvement of 8.3% in the TAP-DLND 1.0 dataset and minimizing mistakes to the extent of 10% for APWSJ). For Baseline 3, instead of multiple premises, we take only the single most relevant premise (having the highest relevance score). It improves over Baseline 2 by a margin

### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

---

of 3.9% for TAP-DLND 1.0 and 5.2% for APWSJ. We observe almost similar behavior for novelty-scoring in TAP-DLND 1.1. However, with our proposed approach, we attain significant performance gain over our ablation baselines, as is evident from our results. Thus our analysis indicates the importance of having *relevance scores* in a *multi-premise* scenario for the task in hand.

#### 3.8.7 Results on Related Tasks

To evaluate the efficacy of our approach, we went ahead to test our model on certain related tasks to textual novelty as indicated earlier.

##### Paraphrase Detection

As already mentioned paraphrase detection is one such task that simulates the notion of non-novelty at the semantic level. Detecting semantic-level redundancies are not straightforward. We are interested in identifying those documents which are lexically distant from the source yet convey the same meaning (thus semantically non-novel). For our purpose, we experiment with the Webis-CPC-11 corpus, which consists of paraphrases from high-level literary texts. We report our results on the paraphrase class as the non-paraphrase instances in this dataset do not conform to novel documents. We perform comparably with our earlier results (Table 3.16). This is particularly encouraging because detecting semantic-level non-novelty is challenging, and the quality of texts in this dataset is richer than simpler newspaper texts.

Table 3.16: Results for paraphrase class on Webis-CPC, 10-*fold* cross-validation output

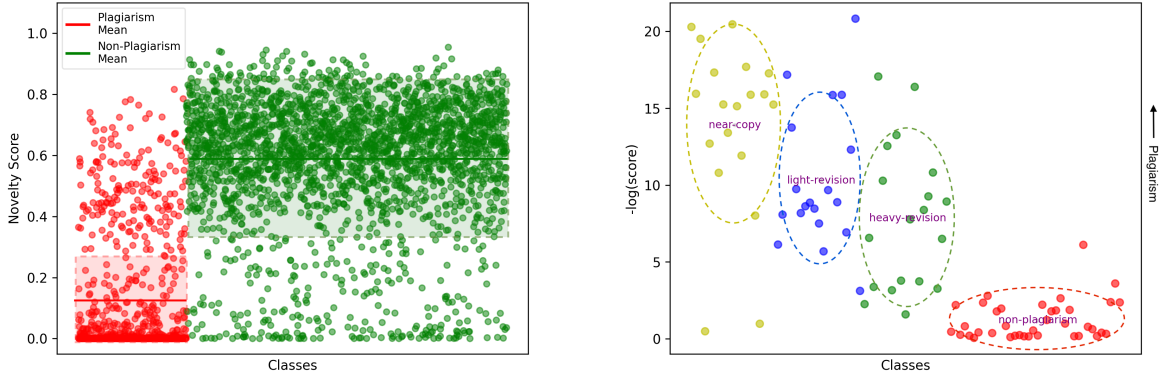
Evaluation System	P	R	$F_1$	A
Baseline 1	0.58	0.69	0.63	58.0%
Baseline 2	0.73	0.92	0.81	77.6%
Baseline 3	0.74	0.92	0.82	78.2%
Comparing System 2	0.75	0.84	0.80	78.0%
<b>Proposed Approach</b>	<b>0.76</b>	<b>0.90</b>	<b>0.82</b>	<b>78.9%</b>

##### Plagiarism Detection

Semantic-level plagiarism is another task that closely simulates non-novelty. The P4PIN dataset is not large (only 847 plagiarism instances) and is not suitable for a deep learning experiment setup. We adapt a *transfer learning* scheme and train our model on TAP-DLND 1.0 (novelty detection task) and test if our model can identify the plagiarism cases in P4PIN. We are not interested in the *non-plagiarism* instances as those do not conform to our idea of *novelty*. Non-plagiarism instances in P4PIN exhibit thematic and stylistic similarity to the content of the

original text. We correctly classify 832 out of 847 plagiarised instances, yielding a **sensitivity of 0.98** towards identifying semantic-level plagiarism. Figure 3.11(a) shows the predicted novelty scores for the documents in P4PIN (trained on TAP-DLND 2.0). We could clearly see that the concentration of novelty scores for the plagiarism class is at the bottom, indicating low novelty, while that for the non-plagiarism class is at the upper half signifying higher novelty scores.

We also check how our model can identify the various degree of rewrites (plagiarism) with the Wikipedia Rewrite dataset. Here again, we train on TAP-DLND 2.0. We take the negative log of the predicted scores (the higher, the less is the novelty score) and plot along the Y-axis in Figure 3.11(b). According to our definition, all the four classes of documents (*near-copy*, *light-revision*, *heavy-revision*, *non-plagiarism*) are not novel. But the degree of non-novelty should be higher for *near copy*, followed by *light revision*, and then *heavy revision*. *Near Copy* simulates a case of lexical-level plagiarism whereas *light revision* and *heavy revision* could be thought of plagiarism at the semantic-level. The novelty scores predicted by our model display the novelty score concentration in clusters for each category. If there is no plagiarism, the novelty score is comparatively higher (non-plagiarism instances are at the bottom signifying higher novelty scores). All these performances of our approach on prediction of the non-novel instances indi-



(a) Novelty scores for plagiarism and non-plagiarism (b) Concentration of novelty-scores for the four classes for P4PIN classes in WikiRewrite Dataset

Figure 3.11: Predicted novelty scores for documents in P4PIN and WikiRewrite by our model trained on TAP-DLND 1.0

cates that finding multiple source and assimilating the corresponding information to arrive at the judgement for novelty/non-novelty is essential.

#### 3.8.8 Analysis

The actual documents in all of our datasets are long, hence, we take the same example in Section 3.8.2 to analyze the performance of our approach.

Figure 3.12 depicts the heatmap of the attention scores between the target and source document sentences. We can clearly see that for target sentence  $t_1$  the most relevant source sentences predicted by our model are  $s_1$ ,  $s_2$ ,  $s_3$ . While we read  $t_1$  (*Facebook was launched in Cambridge*) against the source document, we can understand that  $t_1$  is offering no new information. But in order to do that we need to do a multi-hop reasoning over  $s_1$  (*Facebook*)  $\rightarrow$   $s_2$  (*created in Harvard*)  $\rightarrow$   $s_3$  (*Harvard is in Cambridge*). The other information in  $s_4$  (*Zuckerberg lives in California*) do not contribute for ascertaining  $t_1$  and hence is a distracting information. Our model pays low attention to  $s_4$ .

Similarly, when we consider the next target sentence  $t_2$  (*The founder resides in California*), we understand that  $s_4$  (*Zuckerberg lives in California*),  $s_2$  (*Zuckerberg created Facebook*),  $s_1$  (*Facebook*) are the source sentences which ascertains that  $t_2$  do not have any new information.  $s_3$  (*Harvard is in Cambridge*) finds no relevance to the sentence in concern. Hence our model assigns lowest attention score to  $s_3$  for  $t_2$  signifying that  $s_3$  is a distracting premise.

Finally, our model predicts that the target document in concern is *non-novel* with respect to the source document. The predicted *novelty-score* was 20.59 on a scale of 100.

Let us now take a bit complicated example.

*Source Document 1 (S1): Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment.*

*Source Document 2 (S2): The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. These droplets are too heavy to hang in the air and quickly fall on floors or surfaces. You can be infected by breathing in the virus if you are within close proximity of someone who has COVID-19 or by touching a contaminated surface and then your eyes, nose, or mouth.*

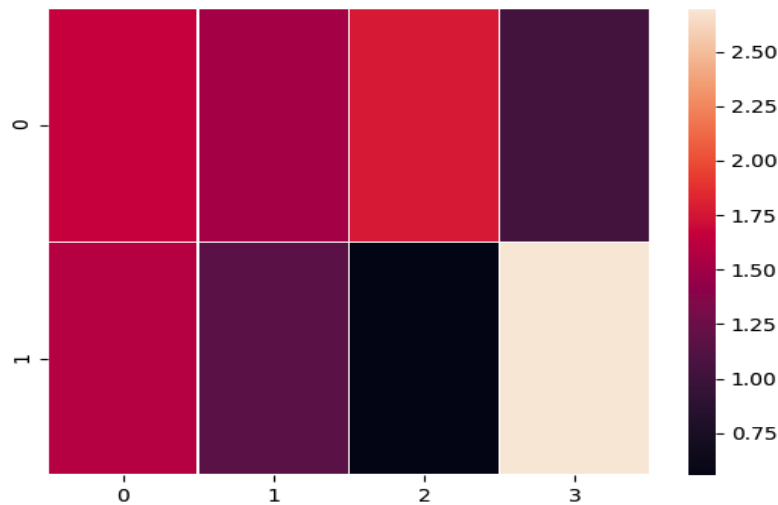


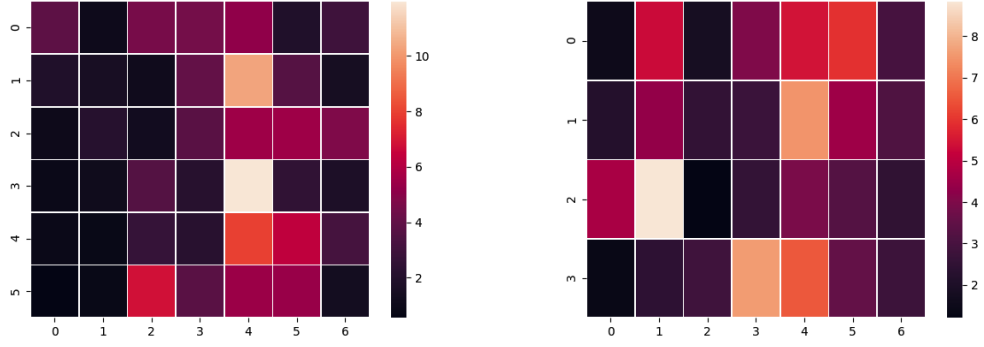
Figure 3.12: Heatmap depicting the attention scores between the source and target document (Example 2 in section 3.8.2).  $t_1, t_2$  are the target document sentences (vertical axes) and  $s_1, s_2, s_3, s_4$  are source document sentences (horizontal axes). The brighter the shade, more is the alignment, signifying an affinity towards non-novelty

*Source Document 3 (S3): You can reduce your chances of being infected or spreading COVID-19 by regularly and thoroughly clean your hands with an alcohol-based hand rub or wash them with soap and water. Washing your hands with soap and water or using alcohol-based hand rub kills viruses that may be on your hands.*

*Target T1 (Non-Novel): Coronavirus is a respiratory illness, meaning it is mostly spread through virus-laden droplets from coughs and sneezes. The government’s advice on Coronavirus asks the public to wash their hands more often and avoid touching their eyes, nose, and mouth. Hands touch many surfaces and can pick up viruses. Once contaminated, hands can transfer the virus to your eyes, nose, or mouth. From there, the virus can enter your body and infect you. You can also catch it directly from the coughs or sneezes of an infected person.*

*Target T2: COVID-19 symptoms are usually mild and begin gradually. Some people become infected but don’t develop any symptoms and don’t feel unwell. Most people (about 80%) recover from the disease without needing special treatment. Older people, and those with underlying medical problems like high blood pressure, heart problems or diabetes, are more likely to develop serious illness.*

### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection



(a) Heatmap for Target T1 against S1, S2, S3 (b) Heatmap for Target T2 against S1, S2, S3

Figure 3.13: Heatmap depicting the attention scores between the source ( $S_1, S_2, S_3$ ) and target document ( $T_1, T_2$ ). The brighter the shade, more is the alignment, signifying an affinity towards non-novelty

The heatmap for the above example after prediction is shown in Figure 3.13. Keeping the source documents ( $S_1, S_2, S_3$ ) same, we analyze our model’s prediction against the two Target Documents ( $T_1$  and  $T_2$ ). The source document sentences are along the horizontal axes, while the target document sentences are along the vertical axes. After reading  $T_1$  and  $T_2$  against  $S_1, S_2, S_3$  we can understand that  $T_1$  is offering very little new information, however  $T_2$  has some amount of new information (*Older people are more susceptible to the disease*). Our model predicts 22.73 and 40.30 as novelty scores for  $T_1$  and  $T_2$  respectively which is somewhat intuitive.

The third sentence in  $T_2$  (*Most people (about 80%) recover from the disease without needing special treatment.*) highly attends the second sentence in  $S_1$  (*Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment*). Similarly, the third sentence in  $S_2$  pays greater attention to the fourth sentence in  $T_1$ , signifying that the target sentence is having less/no new information with respect to the source candidates.

We can see via the above heatmap figures how multiple premises in the source documents are attending the target sentences, which is correctly captured by our approach hence establishing our hypothesis.

#### 3.8.9 Error Analysis

We have identified a few causes of errors committed by our approach.

- **Long Documents:** The misclassified instances in the datasets (APWSJ, TAP-DLND 1.0) are too long. Also, the corresponding source documents have a good amount of

information. Although our architecture works at sentence-level and then composes at the document-level, finding the relevant premises out of large documents is challenging.

- **Non-coherence of Premises:** Another challenge is to aggregate the premises as the premises are not in a coherent order after selection in the Selection Module.
- **Named-Entities:** Let us consider a misclassified instance with respect to the COVID-19 source documents in the earlier example.

*Target T3 (Novel): The world has seen the emergence of a Novel Corona Virus on 31 December 2019, officially referred to as COVID-19. The virus was first isolated from persons with pneumonia in Wuhan city, China. The virus can cause a range of symptoms, ranging from mild illness to pneumonia. Symptoms of the disease are fever, cough, sore throat, and headaches. In severe cases, difficulty in breathing and deaths can occur. There is no specific treatment for people who are sick with Coronavirus, and no vaccine to prevent the disease.*

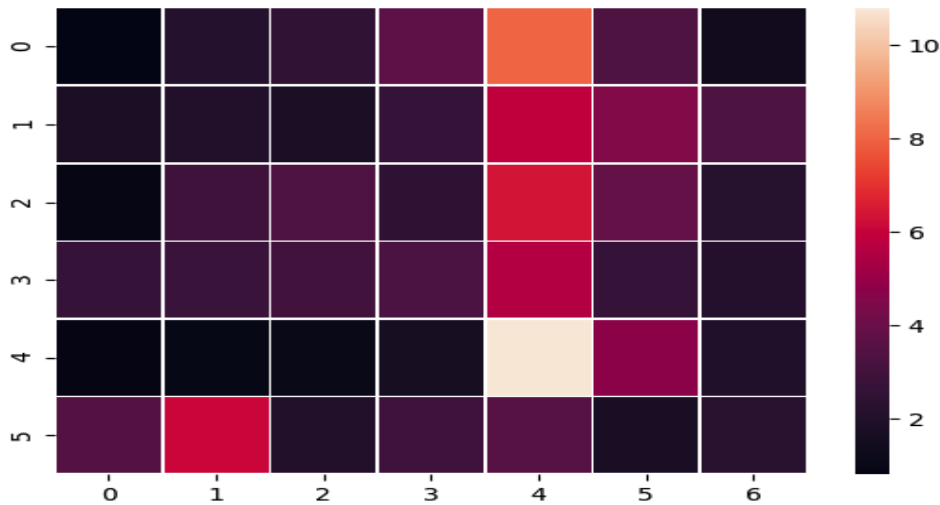


Figure 3.14: Heatmap of the misclassification instance

We could clearly understand that T3 has new information with respect to the source documents. But due to higher correspondence in named-entities and certain content words (e.g., virus) between source-target pair, our classifier may have got confused and predicted T3 as non-novel (Figure 3.14).

Kindly note that our documents in the actual datasets are much longer than the examples we demonstrate, adding more complexity to the task.



### 3.8 Leveraging Multi-Premise Textual Entailment for Novelty Detection

---

#### 3.8.10 Insights gained from the work

A major insight that we gain from this work is that *relevance detection* is a prelude to *novelty detection*. Its not worth investigating novelty if the relevance criteria is not preserved. Our multi-premise entailment-based model testifies that. Also, a single premise is not sufficient to conclude on the novelty of the target sentence, we would need to assimilate information from multiple source sentences to do that. Our RDV-CNN and SETDV-CNN architecture reveals that a semantic composition unit encompassing useful information from both source-target pairs proves crucial to the novelty classification task. The decomposable attention-based architecture is much simple and shows the power of simple alignment between source-target document pairs to find the novelty classification of the target document. *More is the alignment, less is the novelty*. We enforce the alignment for RDV-CNN via cosine similarity whereas via entailment probabilities for our multi-premise entailment-based model. The reason for the better performance of our multi-premise entailment-based model is because of the presence of the relevance detection module to weed out the irrelevant source sentences from the computation. Also, the relationship between entailment and novelty is long-established.

#### 3.8.11 Limitations of our research

The major limitation of our current research would be *semantic compositionality* and *scalability*. Composing semantics from sentential information units is still an active area of research in NLP. For deciding on novelty at the level of documents we need to have a robust semantic representation of premise documents which are currently unavailable. Our models would suffer when there would be a large number of source documents against whom the novelty of the incoming document is to be ascertained. As finding relevant premises is one mandatory clause for our architectures, hence if the search space of the premises/source documents increase, naturally it will impact the performance of our methods. One way to address this could be to do topical clustering of relevant source documents to weed out the outlier documents to reduce the search space. In our further explorations we would look forward to implement this idea and see if that affects our performance.

#### 3.8.12 Transcending to Scientific Novelty - How difficult is the problem?

Finding novelty in scientific articles is fascinating but a very hard problem even for humans. It requires deep human expertise and knowledge to comprehend the novelty of a scientific article which comes through years of experience, study, knowledge assimilation from multiple sources/-domains of knowledge. Scientific knowledge assimilation is not only from texts, hence it is

not that only texts contributes to comprehension of scientific knowledge. Our approaches are solely dependent on background texts which are finite in number. Also our approaches are on *surface-form* newspaper texts which do not hold much background or contextual information. Whereas scientific paper texts are *intelligent* texts and they have background information which are only comprehensible to a knowledgeable reader. Hence our approaches are not tuned for scientific texts. We are yet to ascertain a way to do scientific novelty. Also, another challenge is that scientific papers contain additional artifacts like images, mathematical equations that contribute to the content. Information extraction from PDFs for such items is not straightforward and then assimilating such information with texts is another challenge. It is hard to identify the source knowledge for a given research paper and hence building such a dataset remains a challenge. It might be do-able for certain category of papers (papers which are an extension of a certain work). We attempted this problem of selecting the relevant premises via citation influence detection in our *research lineage* problem which we explore in Chapter 5. It seems logical to attempt relevant premise detection for a given scientific hypothesis as the first step for scientific novelty. We conclude that our present approaches for document-level novelty detection from surface-form newspaper texts would not work well with scientific texts because of the above mentioned reasons.

### 3.8.13 Conclusion

Textual Novelty Detection has an array of use-cases starting from search and retrieval on the web, NLP tasks like plagiarism detection, paraphrase detection, summarization, modeling interestingness, fake news detection, etc. However, less attention is paid to the document-level variant of the problem in comparison to sentence-level novelty detection. In this work, we present a comprehensive account of our experiments so far on *document-level novelty detection*. We study existing literature on textual novelty detection as well as our earlier explorations on the topic. Here we assert that we would need to perform assimilation of information from multiple premises for identifying the novelty of a given text. Our current approach performs better than our earlier approaches. Also, we show that our method could be suitably applied to allied tasks like *Plagiarism Detection* and *Paraphrase Detection*. We point out some limitations of our approach, which we aim to explore next.

In the future, we would aim to explore novelty detection in scientific texts, which would be way more challenging than newspaper texts. We would also like to investigate how we could address situations when the number of source documents would be exponentially higher.

## 3.9 Chapter Summary

In this chapter we discuss our methods for document-level novelty detection. We also discuss the development of our benchmark resource(TAP-DLND 1.0 and TAP-DLND 2.0). Next, we show how we use similarity and diversity-based handcrafted features for the problem and the importance of certain features. Subsequently, we discuss how we encapsulated source and target document information in one unit for feature extraction by a deep CNN for novelty classification as well as quantification. Further we discuss the role of *alignment via neural attention* mechanism to identify novel and non-novel documents. Finally, we explore the role of multi-premise entailment for the problem with our latest deep architecture. We are looking forward to use our methods developed so far with scholarly texts. However, the first step would be *Premise Selection* for a given scientific claim in order to ascertain the novelty of that claim.



## CHAPTER 4

---

# Scope Detection of Research Articles

---

In this second contribution chapter, we explore our investigations towards identifying *out-of-scope* articles in peer review. Our objective is to assist the editors to quickly ascertain if a given submission is within or *out-of-scope* of the venue concerned, thereby saving time of both the authors and the editors. We investigate three different approaches to the problem including handcrafted feature engineering to multimodal deep architecture to multiview-clustering.

---

## 4.1 Introduction

Peer Review is the benchmark of modern-day research validation. In spite of having certain inherent flaws like sometimes being biased, time-consuming, arbitrary [202], peer review is still the widely accepted method to document scientific progress. However, with the exponential rise in article submissions, thanks to the *Publish or Perish* syndrome in academia [203], the peer review system is threatened with a never-seen-before information overload [204]. The electronic preprints repository arXiv receives 500-600 new submissions daily with an additional 300-400 submissions update<sup>1</sup>. Editors and Conference Chairs are overwhelmed with the huge number of submissions made<sup>2</sup>, and they face a dearth of good reviewers to review the submissions [205]. Sometimes the editors and chairs are left with no other option than to assign papers to novice/out-of-domain researchers or graduate students which often results in poor quality reviews, thus affecting the subsequent decision and the entire academia in general. However, studies [67, 206] show that a good number of submissions are not at all informed ones and sometimes are submitted to wrong venues. Unfortunately, in spite of having merit, some articles do not fit to the aims and scope of the intended venue and have to suffer *Desk-Rejection*.

After submission, the first stage in the academic peer review process commences at the editors' desk, wherein the journal editor decides whether the submitted article fits the aims and scope of the concerned journal. The Editor-in-Chief always looks at the scope of the research study with respect to that of the journal before deciding whether to send it for review. Surprisingly a lot many submissions are rejected at the desk [207] popularly known as Desk-Rejection. It means the editor of the particular journal deems the submitted article unsuitable enough to forward to the expert reviewers for meticulous evaluation. Many reasons account for this activity, foremost being that the submitted article is *out-of-scope* of the intended journal [206, 208]. It may signify that the research findings are of interest to a very narrow or specialised audience that the journal does not cater to specifically. A study on the recently released PeerRead dataset [5] reveals that *Appropriateness* of a manuscript to a certain conference (ACL 2017) is the most correlated aspect with the final recommendation by the reviewers<sup>3</sup>.

Our objective here in this work is to reduce this category of information overload and help the editors to identify potential misfit submissions. With the current state of AI, we do not support a fully automated system. Rather we vouch for an editorial assistant who could isolate potential *out-of-scope* submissions to be further looked upon by editors/chairs and thereby speed up the review process. We strive to seek automation that would benefit both the scholars and

---

<sup>1</sup><https://blogs.cornell.edu/arxiv/2018/01/19/a-day-in-the-life-of-the-arxiv-admin-team/>

<sup>2</sup>Apparently CVPR, NIPS, AAAI 2019 received over 5100, 4900, 7000 submissions respectively!

<sup>3</sup>a positive correlation of as high as 0.49

## 4.2 Problem Definition

---

editors to judge the appropriateness of a specific scientific article to the scope of the prospective journal and thereby assist them in making intuitive decisions.

The motivation behind this work is to efficiently manage the exponential rise in article submissions to journals and conferences these days [209]. The rapid growth in scientific production may threaten the capacity for the scientific community to handle the ever-increasing demand for peer review of scientific publications [205]. In spite of having merit, many papers ( $\sim 30\%$ ) [67, 206] are rejected from the desk simply because they are a misfit to the journals aims, scope, and audience. However unfortunate it is, this phenomena still consume the precious time of all the stakeholders (authors/editors/program chairs and even sometimes reviewers) associated in the peer review pipeline. Thus a system of this kind could eventually assist the journal editors and conference chairs to make better-informed decisions regarding the appropriateness of an article to a submitted venue and quickly locate inappropriate *out-of-scope* submissions. Even potential early-career authors may reap the benefit, and they could be confident about the aptness of their research to the desired journal/conference. This would prove as a huge time-saver for both authors and editors and eventually speed up the overall peer review process.

## 4.2 Problem Definition

We frame our investigation as a binary classification problem of research articles (IS: *in-scope* and OS: *out-of-scope* classes). Given a journal  $J$  and a paper  $P$ , we seek to answer: *If  $P$  is within the scope of  $J$ .* Articles already published signify that they are within-the-scope and somewhat define the *domain-of-operation* of a journal. Our *out-of-scope* data are those desk-rejected manuscripts which according to the editors are not a good fit to the topical coverage and aspirations of the journal. To the best of our knowledge, this work is the first attempt towards automatic scope detection for peer review.

## 4.3 Scope Detection

Submitting a manuscript to an unsuitable journal is one of the most common mistakes committed by authors. Usually, novice/early-career researchers and sometimes even seasoned researchers commit this error. The *scope of a journal* is a very broad term and vary across different journals<sup>4</sup>. We enlist some of our observations from the study of Desk-Rejected due to Out-Of-Scope (DR-OOS) articles. Special thanks to our academic collaborator Elsevier, to support this investigation with necessary resources.

---

<sup>4</sup><https://wordvice.com/choosing-the-right-journal-scope-issues/>

### 4.3.1 Desk-Rejection Observations

- If one submits a paper from Computer Networks to an Artificial Intelligence journal; it is out-of-scope. However, naive as it may sound, this activity not rare, ultimately resulting in desk-rejection.
- Again a paper which is too specific to a particular domain of interest (e.g., Neural Networks) sometimes is not accepted by a journal which caters to a broader perspective (e.g., Artificial Intelligence).
- Similarly, a journal which accepts review papers (e.g., ACM Computing Surveys) may not consider a method paper and vice-versa.
- A theoretical journal (e.g., Theoretical Computer Science) would not be interested in an application-focused paper even though the domain may be identical.
- Sometimes the scope is also linked to the quality of the manuscript. A journal may cater to a vast area of topics but only looks for high quality, original and innovative submissions (for example Nature or Science) which have the potential to induce a significant impact post-publication.
- Again we observe that *scope* of a journal is *time-variant* and usually gets streamlined over time. This behaviour reflects the advancements in science and popularity of topics in the scientific community (for e.g., Deep Learning is hugely popular now in NLP, AI, and CV community).

Most of the journals ask the potential authors to go through the past accepted papers of that journal to get a feel of the type of papers they publish and the audience they cater to. The past publishing activity of a journal defines its *domain of operation* and the topics it is interested in. However, the problem of misinformed submissions is still glaring at the present-day peer review system; authors do make less-informed choices, resulting in wastage of precious time of both the authors and journal editors.

### 4.3.2 Scope of a Journal

Scope, simply stated, is the journal's purpose or objective. It is what the publication wants to achieve by delivering its content to the readers. The relevance/similarity of an article with published papers is a good indicator of its domain. However, the article should not be that similar so that it falls short of the originality/novelty criteria. The domain of a journal is one



## 4.4 Feature-based Machine Learning for Scope Detection

---

variant of its scope. In spite of having merit, many submissions face rejections because they do not fit to the declared domains of the journal. So we understand that *Scope* of a journal is very subjective and is hard to define in quantitative terms. There are many views, and we could aptly cast it as a multiview problem. However, in this work, we attempt to explore a limited definition of journal scope: *the domain or range of topics a given journal caters to*. Our experience with the study of desk-rejected papers reveals that out-of-domain submissions are common and account for a large number of desk rejections. Here we try to understand which section of the manuscript contributes more to define its domain and belongingness to a particular journal. However, this in no way mitigates the broader perspective of scope we discussed earlier. Even the available journal recommender systems check the *domainness* of a manuscript to its published articles by simple content words match. *Accepted published articles are thus the benchmark of reference*.

## 4.4 Feature-based Machine Learning for Scope Detection

We extract features from almost every section of a scientific manuscript that could contribute to identifying its domain: *Author, Content and Bibliography*. Our point of departure for this particular work is the bibliography section of research articles. We intuitively hypothesize that with obvious exceptions *if an article belongs to a particular domain then the majority of its references would fall in that certain domain*. Coupled with other factors, our approach *ScopeJr* achieves *state-of-the-art* performance across six different journals. However, we agree that this preliminary work may not hold universally for all journals as the nature of the scope of different journals is different. Especially for interdisciplinary journals, journals having a wider domain of operation like *Nature, Science*, etc., our hypothesis won't hold. *Scope* and *Quality* of the article would have overlapping requirement criteria in these cases.

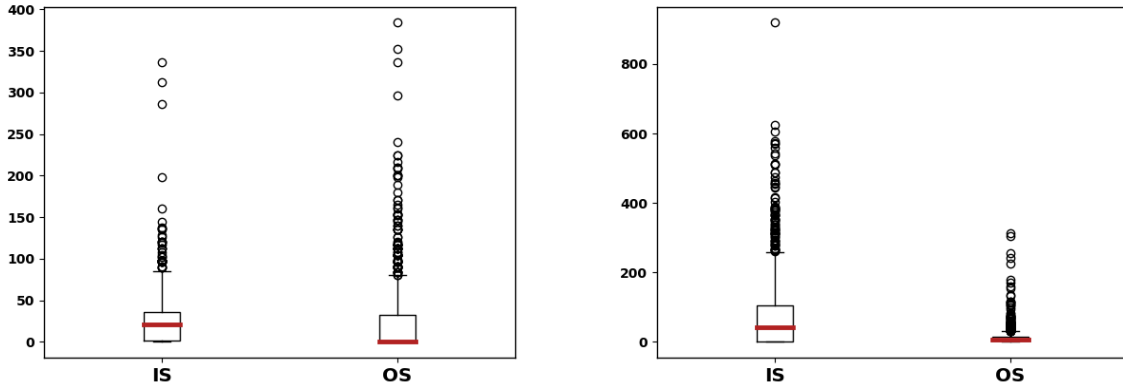
### 4.4.1 Data Description and Preprocessing

We consider all accepted (ACC) articles from six different Elsevier Computer Science journals: Artificial Intelligence (ARTINT), Computer Networks (COMNET), Statistics and Probability Letters (STATPRO), Theoretical Computer Science (TCS), Computer Standards and Interfaces (CSI), and Simulation Modeling Practice and Theory (SIMPAT), to build our domain lists. However, for our experiments, we use 1000 exclusive accepted articles from each journal as our *In-Scope* data. We procure and curate 1000 *out-of-scope* articles for each of these journals internally from Elsevier (most of them were actually desk-rejected due to out-of-scope, and some were accepted articles of distantly related journals<sup>5</sup>). After a thorough study of these data, we

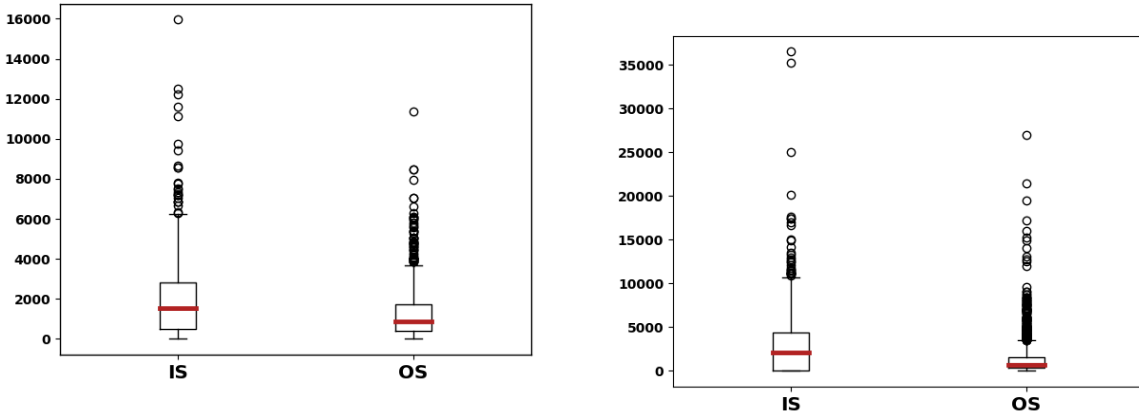
---

<sup>5</sup>Rejected articles are confidential and hard to get

come up with some important observations (Fig 4.1). There are noticeable differences among In-Scope (IS) and Out-of-Scope (OS) articles in terms of the keywords they use, bibliography (papers and venues) they refer and their authors when compared to corresponding past accepted papers. For e.g., the median of bibliographic title overlap for IS articles against ARTINT-ACC articles is found to be 43 whereas the same for OS articles is 7 (Figure 4.1b). We observe similar contrast in data distribution across IS and OS articles for various other factors (venue, keyword, authors) as well. Hence we curate our features based on these observations. We parse<sup>6</sup>



(a) Keyword overlap with past ACC data (STAT-PRO) (b) Referenced paper overlap with past ACC data (ARTINT)



(c) Referenced venue overlap with past ACC data (COMNET) (d) Author overlap with past ACC data (ARTINT)

Figure 4.1: Box plots of various factors across an exclusive set of 1000 IS and 1000 OS articles. The match is in terms of overlap of keywords, referenced paper titles, bibliographic venues and authors with respect to past accepted papers of each journal. The median is always high for IS w.r.t. OS articles.

the scientific articles, originally in .pdf format, to generate a corresponding .xml document

<sup>6</sup>using GROBID: <https://github.com/kermitt2/grobid>

## 4.4 Feature-based Machine Learning for Scope Detection

---

consisting of essential information within structured XML tags. We then extract the following information from these .xml versions: *Title, Author names, Abstract, Author-listed keywords, Body-text, Bibliographic Paper Titles and corresponding Venues*. The extracted data are noisy, and we perform certain **pre-processing**:

- Removed editions from conference names and mapped different editions of the same conference and abbreviations into one. For e.g., *Proceedings of the 8th ACM International Conference on Mobile Computing and Networking* → ACM International Conference on Computer and Communications Security → CCS
- Mapped variants of certain words in conference or journal names via regular expressions. For e.g. *Jour.* → *Journal*, *Trans.* → *Transactions*, *Distrib.* → *Distributed*

### 4.4.2 Methodology

We briefly describe our method here. The first task is to build several domain lists from several sections of the paper.

#### Building Domain Lists

As the past accepted articles are strong indicators of the *domain of operation*, or *scope* of a particular journal, we build our features based on the information extracted from those published ones. We pay special attention to the *bibliography* section. For each journal, we create several exhaustive lists:

- (L1) A **Keyword Dictionary** consisting of *author-listed* keywords and record their frequency of occurrences (average 5k+ keywords per journal)
- (L2) **Bibliographic Title List** (average 30k+ titles per journal)
- (L3) **Bibliographic Venue List** (average 7k+ venues per journal)

We hypothesise that for a particular journal, *the relative importance of some papers and certain venues would always be high if measured across all published articles*. Those certain papers/venues are the representative data points of that journal. Hence we extract all referenced paper titles and corresponding venues from the bibliography section of ACC articles. For each entry  $X$  in **L2, L3**:

$$V(X) = \sum_{j=1}^n \text{CitE}(X) \quad (4.1)$$

where  $X$  could either be a **paper title** or a **venue** (journal or meeting/conference/workshop) in the reference section of article  $j$ .  $n$  is the number of ACC articles. We define a novel function **Citation Effect** ( $CitE$ ) which:

- corresponds to the number of in-citations of  $X$  within the body of a candidate article  $j$  if  $X$  is a **paper title**.
- corresponds to the number of occurrences of  $X$  within the bibliography section of article  $j$  if  $X$  is a **venue**.

- (L4) **Author List** (average 15k+ authors per journal)

The intuitions behind creating such lists are:

- Articles which are highly in-cited within a particular journal have higher relevance to the scope of the journal.
- Similarly, venues which have a higher presence in the bibliography section of a particular journal, are of higher relevance to the scope of that journal.
- More an author publishes articles belonging to a certain domain; greater is the chance that her prospective next would belong to the same domain (authors' favourite). We record the publication frequency of authors in each journal separately.

## Feature Engineering

We describe our curated features in this section.

1. **Weighted Keyword Match[wt\_kw\_m]**: We design this feature to emphasise the containment and relative importance of the keywords in the candidate article with respect to the Keyword Dictionary. The value for this feature for a candidate article  $Y$  is:

$$KWScore_Y = \frac{|KW_Y \cap KW_D|}{|KW_Y|} \times \sum_{i=1}^{|KW_Y \cap KW_D|} f(K_i) \quad (4.2)$$

where  $KW_Y$  is the set of author-defined keywords in the candidate article  $Y$ ,  $KW_D$  is the set of keywords in the Keyword Dictionary  $D$ ,  $f(K_i)$  is the frequency of keyword  $K_i$  as listed in  $D$ , and  $K_i \in \{KW_Y \cap KW_D\}$ . Frequently occurring keywords are domain-specific words, hence have higher weights.

2. **Title Scope** and **Venue Scope**: We calculate these features from the bibliography section of a candidate article  $Y$ . From the two exhaustive lists of paper titles and venues, we

#### 4.4 Feature-based Machine Learning for Scope Detection

---

calculate the Title Scope ( $T_Y$ ) and Venue Scope ( $V_Y$ ) respectively:

$$T_Y = \sum_{k=1}^m [V(t_k) * CitE(t_k)] \quad V_Y = \sum_{k=1}^m V(v_k) \quad (4.3)$$

where  $m$  is the total number of bibliographical references in  $Y$  ;  $V(t_k)$  is derived from table look-up Bib-Title List and  $CitE(t_k)$  is the citation effect of  $k$ -th title in  $Y$ . Similarly  $V(v_k)$  is derived from table look-up Bib-Venue List.[**bib\_tit\_sc**, **bib\_jr\_sc**, **bib\_cnf\_sc**]

3. **Author Domain Publication Frequency[adpf]** For a candidate article, we take the summation of the publication frequency of its authors in the concerned journal from the author list.
4. **Distance From Cluster of Similar Articles[clust\_dist]** The accepted articles of a specific journal are grouped into clusters representing different sub-domains within the journal scope. Thus the distance of a given research article from the set of clusters formed on the accepted articles may contribute to determine its scope. Any outlier to such clusters may be considered as *out-of-scope*. With this intuition we perform the steps in Algorithm 1. For each journal, we generate clusters from all accepted articles. We then take *minimum* of the distances of the candidate article  $Y$  from the cluster centers, in order to learn how close is  $Y$  to any of the clusters so formed.

#### Computer Science Specific Word Embeddings

*One major contribution in executing this feature is the creation and usage of **word2vec**[163] word vectors trained on the entire Computer Science journal articles of Elsevier (to preserve domain dependency). We processed 41737169 sentences from around 400K articles. The embedding dimension is set to 300. We choose lines of texts extracted from *Title*, *Abstract*, *Introduction*, *Body*, *Conclusions* sections of accepted articles. Certain preprocessing needs are: removal of special characters, headings, table and figure captions, etc.*

#### 4.4.3 Evaluation

To evaluate the performance of our system we employ a range of classifiers on our feature set. However due to the inter-dependent nature of our features we find that Random Forest performs the best across all journals. We coin our approach using Random Forest classifier as **ScopeJr**. For each of the journals we take 1000 exclusive accepted papers as *in-scope* data and 1000

---

**Algorithm 1** Calculate distance from journal cluster boundary

---

- 1: Use RAKE[210] to automatically extract keywords from the Title, Abstract, Introduction and Conclusion sections of an article  $Y$  belonging to journal  $J$ .
- 2: Use *word2vec*[163] to generate the vectors of the extracted keywords (top 30 ranked RAKE extracted keywords) from  $Y$ .
- 3: Calculate the document vector of  $Y$  by concatenating all the keyword vectors from *Step 2*.
- 4: Repeat *Steps 1-3* for all the accepted articles of the journal  $J$ .
- 5: Use Word Mover's Distance (WMD)[16] as the distance metric between two document vectors and generate the similarity matrix.
- 6: Apply K-Medoids[211] on the similarity matrix from *Step 4* to generate the clusters ( $C_i$ ) [ $K$  is determined via Silhouette Index; user tune-able; can vary across journals]
- 7: Find the radius( $r_i$ ) of a cluster  $C_i$  as:

$$r_i = \text{median}(\text{distance}(c_i, p_j))$$

where  $c_i$  is the centre of cluster  $C_i$  and  $p_j$  is any point within cluster  $C_i$ .

- 8: Find the document vector ( $p_Y$ ) of a candidate article  $Y$  using *Steps 1-3*.
- 9: Distance of the candidate article  $Y$  from the boundary of cluster  $C_i$  is given as :

$$D_i = \text{distance}(c_i, p_Y) - r_i$$

- 10: Repeat *Step 9* for all the clusters ( $C_i$ ) obtained from *Step 6* to get :

$$D_Y = \text{minimum}(D_i)$$


---

*out-of-scope* articles as rejected data. We extract features and perform the experiments in a *10-fold cross-validation* classification set up. Finally we compare the classification performance of our proposed system with the *state-of-the-art* **Elsevier Journal Finder (EJF)**[103] on the same dataset and report the results. EJF is a *state-of-the-art* recommender system provided by Elsevier solutions to the academic fraternity which recommends highly relevant journals to the authors for their papers. Elsevier Journal Finder takes as input the *Title* and *Abstract* of a prospective scientific article ( $Y$ ) and presents a list of 10 relevant Elsevier journals ( $J$ ) to the user as output which s/he may consider for submitting her/his article. Although the recommended journals are limited only to Elsevier published ones, but it is to be noted that Elsevier has more than 2900 peer-reviewed journals that cover almost all major scientific domains. Although we had *true class* labels from Elsevier data, we follow heuristics to determine the **EJF predicted** class label of a prospective article  $Y$  : *If EJF suggests  $J$  for  $Y \rightarrow Y$  is **In-Scope** of  $J$  otherwise, EJF deems  $Y$  to be **Out-of-Scope** for  $J$ .*

**Baseline:** We take the weighted overlap of keywords extracted from *Title*, *Abstract* with Keyword Dictionary (D) as features. We use standard Support Vector Machine (SVM) as the classifier.

#### 4.4 Feature-based Machine Learning for Scope Detection

Table 4.1: Scope-Check figures for *out-of-scope* (OS) class across 6 journals,  $P \rightarrow Precision$ ,  $R \rightarrow Recall$ ,  $\ddagger \rightarrow$  Baseline using only Title and Abstract with SVM classifier. The Accuracy values ( $\dagger$ ) for *ScopeJr* are statistically significant over EJF performance (two-tailed t-test,  $p < 0.05$ )

Journals→	ARTINT			COMNET			STATPRO		
Approaches↓	P	R	A	P	R	A	P	R	A
Title+Abstract $\ddagger$	0.49	0.58	55.8	0.56	0.64	58.2	0.44	0.49	48.9
EJF	0.54	0.62	63.6	0.34	0.43	44.4	0.43	0.52	53.5
<i>ScopeJr</i>	0.89	0.86	<b>87.2<math>\dagger</math></b>	0.82	0.80	<b>81.4<math>\dagger</math></b>	0.83	0.84	<b>83.9<math>\dagger</math></b>
Journals→	TCS			CSI			SIMPAT		
Approaches ↓	P	R	A	P	R	A	P	R	A
Title+Abstract $\ddagger$	0.43	0.45	46.4	0.49	0.58	61.2	0.54	0.63	63.2
EJF	0.55	0.64	66.8	0.51	0.67	65.6	0.53	0.65	64.8
<i>ScopeJr</i>	0.86	0.87	<b>87.2<math>\dagger</math></b>	0.81	0.95	<b>86.7 <math>\dagger</math></b>	0.72	0.76	<b>72.2<math>\dagger</math></b>

##### 4.4.4 Results and Observations

Results reported in Table 4.1 demonstrate the richness of our feature set. Using our feature set with Random Forest (RF), our approach *ScopeJr* performs way better than the baseline and Elsevier Journal Finder (EJF). Except for in SIMPAT, we achieve an improvement of over 20% in terms of accuracy. The comparatively low performance in SIMPAT is because SIMPAT has a wider scope, mostly simulations of different theories, and accepts articles from different disciplines. Since we are particularly interested in a pruning perspective, we report *out-of-scope* (OS) results. Thorough analysis of data and experimental results led us to the following observations:

1. *Bibliographic* features have induced significant improvements (Figure 4.2) because we deduce the *Bibliographic feature* values from within the body section of the scientific articles. *When a certain portion of a scientific article cites a reference, the scope of that portion is influenced by the domain of referenced article. The domain of the cited reference exerts local influence on that portion of the scientific article.* So if many in-domain references are cited in distributed portions of a research article, quite possibly the entire research article falls in the same domain. We measure *in-domain* or *in-scope* by simply counting occurrences across published articles of a certain journal; higher the better.
2. For all the journals our approach outperforms the **EJF** in terms of precision, recall and accuracy values. This is due to the fact that EJF considers only the *Title* and *Abstract* sections of a research article and uses the **Elsevier Finger Print Engine**<sup>7</sup> based on identification of *Noun Phrases* from those sections. Our method goes beyond this idea and we use *Bibliographic*, *Author* and *Content* information which highly contributes the towards categorization.

<sup>7</sup><https://www.elsevier.com/solutions/elsevier-fingerprint-engine>

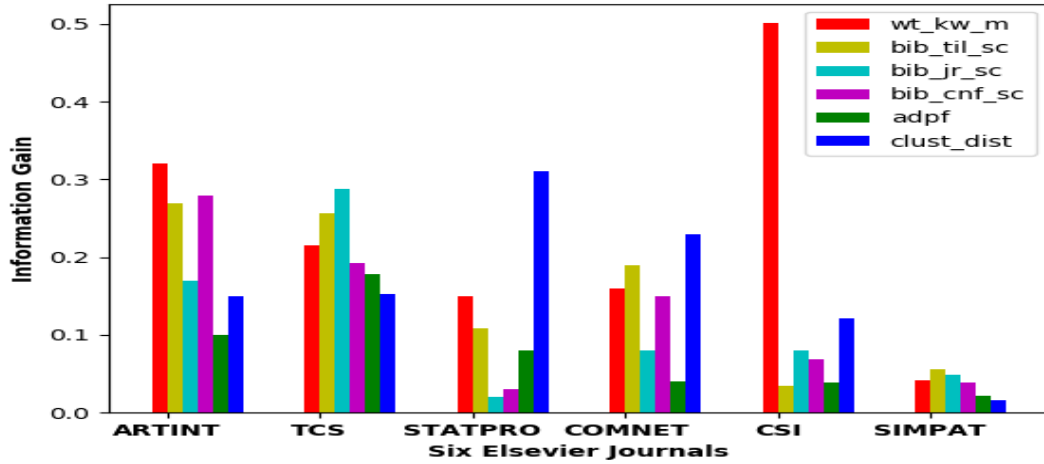


Figure 4.2: Significance of features observed by ranking features based on Information Gain

3. Some journal specific features (like presence of mathematical expressions for STATPRO) may further improvise performance.
4. For journals having very wider scope (for e.g., Computer Science Review or Nature or Science) or multi-disciplinary in nature, this approach may not be fruitful.
5. Scope of a journal gets more compact and streamlined with time. Hence experimenting with only recent articles instead of historical ones may boost the performance.
6. Journals SIMPAT and CSI accept papers across many domains. Hence we observe information from several domains in their *Bibliography* section. However for ARTINT, STATPRO, and COMNET we find *Bibliography* generates a comparatively restricted *domain-specific* set and hence bibliographic features proved more effective.
7. Some authors co-author multiple publications in the same journal which signifies their area of interests. New authors usually have supervisors as co-authors; hence we do a summation of the frequency of their publications. We see *adpf* feature has less significance in comparison to others. However, this feature could be important if we consider an entire domain consisting of different journals as the reference list.



## 4.5 Multimodal Scope Detection

Research articles are essentially multimodal, especially considering those from STEM disciplines. The variety of figures, graphs complements the text in the article and enables the reader to understand the proposition and analysis better. While images may not always be that significant to certain disciplines, but do play a major role in the comprehension of the research in others (for e.g., natural sciences and medicine [212]). Here in this work, we are intrigued to see if images in research articles contribute to this problem of domain-based research article classification. Our objective is to make use of every possible channel of information in a research article for scope classification. We design a multimodal deep neural architecture and investigate the role of full-text, bibliography, embedded images to determine its appropriateness to the concerned venue. Our approach does not involve any handcrafted features, solely depends on the past accepting activity of the venue, and thereby achieves significant performance on two real-life datasets. Our findings suggest that a system of this kind is possible and with reasonable accuracy could assist the editors/chairs in flagging out inappropriate submissions. To the best of our knowledge, this work is the first to explore multimodal information from scholarly papers for scope detection.

### 4.5.1 Problem Definition

Given a set of  $N$  research articles, the objective is to minimize the negative log likelihood over the classes:

$$-\sum_{c=1}^M y_{o,c} \log(p_{o,c})$$

where  $p_{o,c} = f(o(x_n))$ ,  $o$  is the network's output,  $x_n$  is the multi-channel multi-modal input and  $y_{o,c}$  is the indicator if class label  $c$  is the correct classification for observation  $o$ . In our case,  $M = 2$  and  $f = \textit{sigmoid}$  for Task 1 (binary classification) and  $M = 3$  and  $f = \textit{softmax}$  for Task 2 (multi-class classification). Here the modality is two: text and image. Further the text modality has two distinct channels: full text and bibliography.

### Task 1

We model the problem as a binary classification one: *classifying a given article into **within-scope** or **out-of-scope** classes*. We train separate models on accepted and *out-of-scope* articles of each venue to test the suitability of an incoming article to the scope of the particular journal/conference.

## Task 2

Here we are interested to see if we can predict the actual venue of an article among potential venues with a nearly identical domain of operation. With this motivation, we design our second set of experiments on Dataset-II. With Dataset-II we want to go deeper and see how the various channels of information contribute to identifying the class of a research article which may belong to multiple venues having overlapping nature of scope. We model the problem to handle multi-class scenarios where the objective would be: *To which venue a particular paper should go when there are multiple potential venues?* We seek how past accepted papers in these venues having overlapping nature of scope could effectively identify the place holder of a new submission? This could be effectively seen from the viewpoint of a prospective author and towards a venue recommender system.

### 4.5.2 Data Description and Analysis

Getting hold of actual desk-rejected data is hard due to proprietary and confidentiality reasons. However, we curate two datasets to proceed with our experiments. One we got from our collaborator (Dataset-I) and the other we create from open access articles (Dataset-II). The motivation behind experimenting with these two datasets are slightly different. While with Dataset-I we cater to the need at the editors' end, with Dataset-II we find an author perspective to the problem. We follow an 80%: 20% split with our data for training and testing respectively. We nearly equate the positive and negative instances in the train/test split.

#### Dataset-I

We create Dataset-I with the papers from the following six Computer Science journals: Artificial Intelligence (ARTINT), Computer Networks (COMNET), Journal of Computer Network and Applications (JNCA), Computer Standards and Interfaces (CSI), Simulation Modelling Practice and Theory (SIMPAT), Statistics and Probability Letters (STATPRO). We are thankful to our collaborator Elsevier, for providing us with a subset of desk-rejected data of these six journals. In our earlier study [208], we show that nearly 50% of desk-rejections accounts for articles not being within scope. However, for a deep learning experimental setup, the actual available *out-of-scope* papers were not sufficient. Hence, along with actual *out-of-scope* instances from the desk-rejected articles, we also select articles from other journals to serve as the negative instances for a given journal. The intuition is simple: *Accepted articles of other remotely related journals would be out-of-scope of the current journal under study.* We consider accepted papers from a set of 17 different Computer Science journals to simulate our negative data. This we do to make

## 4.5 Multimodal Scope Detection

Table 4.2: Dataset-I Statistics (Elsevier), FT→Full-Text, Actual Negative are the instances (papers) which were desk-rejected due to *out-of-scope* from the concerned journal, Bib→Bibliography, J1→ARTINT, J2→COMNET, J3→STATPRO, J4→JNCA, J5→SIMPAT, J6→CSI, Tr→Training Data, Tt→Test Data

J	Accepted		Rejected		Actual Negative		#Images		#Bib. Entries		# FT Sentences	
	# Tr	# Tt	# Tr	# Tt	# Tr	# Tt	# Tr	# Tt	# Tr	# Tt	# Tr	# Tt
<b>J1</b>	1348	337	1239	308	270	68	5660	1518	44379	11174	725791	174611
<b>J2</b>	2957	740	2934	729	365	92	6878	2685	64613	16602	1017532	261342
<b>J3</b>	4345	1087	3956	981	307	77	1734	910	25598	6725	646893	160351
<b>J4</b>	1614	404	1450	365	24	6	14923	3705	45958	10629	938470	234754
<b>J5</b>	1228	307	1149	285	419	103	7850	4093	24454	6222	325053	86010
<b>J6</b>	1663	416	1499	375	17	5	4532	1303	16700	4748	287150	76769

our negative data as diverse as possible. We had all the accepted articles of the six journals as our positive data. Table 4.2 illustrates our Dataset-I statistics. The total number of images and bibliography items for each journal speaks high of the volume of information they carry within the manuscripts. ARTINT journal invites original research in theory, techniques and applications of Artificial Intelligence. The domain is vast. COMNET is for topics on Computer Networks and is somewhat restricted in the area as compared to ARTINT. JNCA is close to the scope of COMNET and has overlapping *topics of interest*. The journal Simulation Modelling Practice and Theory (SIMPAT) provides a forum for original, high-quality papers dealing with any aspect of systems simulation and modelling. Computer Standards and Interfaces (CSI) focusses on quality of software, well-defined interfaces (hardware and software), the process of digitalisation, and accepted standards in these fields. STATPRO is all about Statistical and Probability theories and has a limited scope as compared to others.

### Dataset-II

Dataset-II comprises of open access articles from several top-tier conferences in the field of Artificial Intelligence and Machine Learning, Natural Language Processing, and Computer Vision. NLP and CV are sub-fields of AI and are currently heavily reliant on ML techniques. Hence there is an overlapping domain of interest between AI and NLP/CV conferences. AI conferences accept papers that address challenges in both NLP and CV. However, AI conferences also cater to several other areas like Robotics, Data Mining, Knowledge Discovery, Machine Learning, etc. With the recent interest and rapid progress in AI/ML domain, every other STEM discipline is using AI/ML, thus making the scope of AI very broad. However, there are some subtle distinctions in aims and motivations behind general AI conferences and more specific venues from NLP and CV. Certain domain-specific papers in NLP and CV would be of more interest to a specialist audience than a general one. Hence with this dataset, we explore, to which conference category a particular paper should belong? With our earlier dataset, the distinctions were

Table 4.3: Dataset-II Statistics (Open Access AI/ML/NLP/CV Papers), This statistics signify the volume of information processing corresponding to the three modalities

Category	Conferences	#Images		#Bibliography		#Sentences		#Papers	
		# Tr	# Tt	# Tr	# Tt	# Tr	# Tt	# Tr	# Tt
AI/ML	IJCAI, AAAI, ICLR, ICML, NIPS	6596	3169	163642	29011	1324259	200424	6719	932
CV	CVPR, ICCV, ECCV	7290	4804	191943	41413	1209511	223876	5403	1011
NLP/CL	ACL, NAACL, EACL, COLING, CoNLL, EMNLP	15200	2666	165345	29456	1193096	190613	5842	920

pretty obvious while with Dataset-II certainly there is an overlap in the *domain of operation* of the venues. We investigate how our deep network trained on previously accepted papers of those allied venues could correctly identify the suitable venue of a new submission. Although we perform a 3-class classification, our model could be suitably tuned to handle multi-class scenarios. The distal objective is to build a recommender system which could efficiently guide the authors to consider a more suitable venue for their manuscripts. We also explore the effects of different modalities in different categories (NLP, CV) of the same domain (AI). Table 4.3 shows the data statistics for Dataset-II. For AI/ML, we consider papers from International Conference on Learning Representations (ICLR), Association for the Advancement of Artificial Intelligence (AAAI) Conference on AI, International Joint Conference on Artificial Intelligence (IJCAI), and NeurIPS (Conference on Neural Information and Processing Systems, previously called NIPS). For NLP, we take papers from Association for Computational Linguistics (ACL), North American Association for Computational Linguistics (NAACL), European Association for Computational Linguistics (EACL), Conference on Empirical Methods in Natural Language Processing (EMNLP), International Conference on Computational Linguistics (COLING), and Conference on Natural Language Learning (CoNLL). For Computer Vision, we consider papers from The Conference on Computer Vision and Pattern Recognition (CVPR), European Conference on Computer Vision (ECCV), and International Conference on Computer Vision (ICCV). Since the rapid progress of Deep Learning in AI has its origin in the ImageNet competition in 2012 [213], we consider papers from these venues from 2012 till 2018.

### Pre-processing

The original articles are in PDF. We use the Science Parse library <sup>8</sup> to convert the PDF into. JSON encoded files for information extraction. Tables, Formulas are distorted in the process, and we exclude those from further processing. We extract figures from the raw PDF's using the PDFFigures 2.0 library [214]. We extract the bibliography section and consider only the citation titles and venues in our experiments. Paper titles and venues contain certain domain-specific

<sup>8</sup><https://github.com/allenai/science-parse>

## 4.5 Multimodal Scope Detection

vocabulary and are a good indicator of the domain of the paper [206]. The other elements in the bibliography (Authors, Year, Page Numbers, Publisher, etc.) has little relevance to our task, and so we ignore them. We remove stop words and certain common words (for e.g., *International*, *Journal*, *Conference*, *Proceedings*, etc.) from the citations. We create a vocabulary list from citation titles and venues and use it in the Bag-of-Words (BoW) model.

### 4.5.3 Methodology

We choose to investigate a deep neural solution to this problem because the definition of scope is not invariant across journals/conferences. Our idea is to let the network learn the *scope* of a venue from its past accepted articles. We present the overall architecture in Figure 4.3. Our architecture is divided into two phases. In Phase-I we learn the feature representation from various modalities. In Phase-II we learn the importance of the modalities via attention mechanism, weigh them accordingly, fuse them, and finally classify the article into *Within Scope* or *Out-of-Scope*.

### Phase I: Representation Learning of Multimodal Paper Features

Here we learn useful features from different paper components (Full-Text, Images, Bibliography).

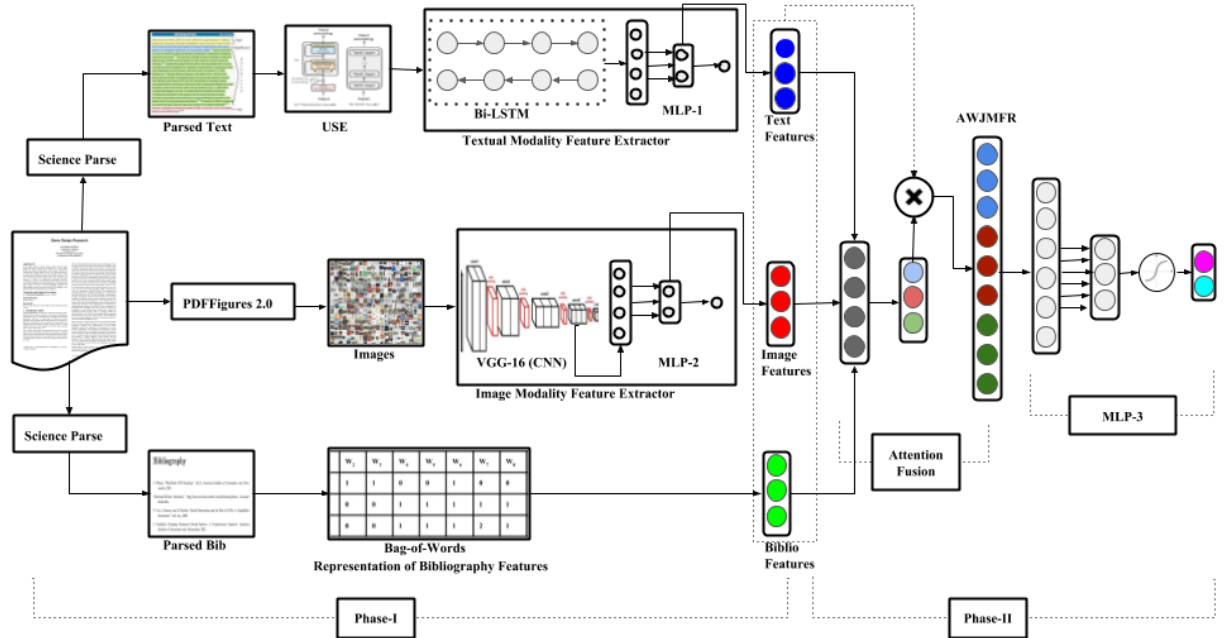


Figure 4.3: Proposed Deep Multimodal Neural Architecture for Scope Detection

### Textual Feature Extraction

We extract full-text sentences from each research article and use the Transformer variant of the Universal Sentence Encoder (USE) [198] to encode the full-text sentences into 512 dimensional semantic vectors. We then stack the sentence vectors to form the document representation. Next we train our *Textual Modality Feature Extractor* by passing this document representation through an end-to-end Bi-Directional Long Short Term Memory (Bi-LSTM) network followed by a Multi-Layer Perceptron (MLP-1) with a final sigmoid layer for classification. We use the activations of the preceding fully-connected layer of MLP-1 as the document-level feature representation of Text T.

Let  $[S_i]$  be the output sentence representation of the Universal Sentence Encoder. We use separate LSTM modules to produce forward and backward hidden vectors, which are then concatenated:

$$\begin{cases} \vec{h}_t = \overrightarrow{LSTM}_t([S_i]) & (4.4) \\ \overleftarrow{h}_t = \overleftarrow{LSTM}_t([S_i]) & (4.5) \\ h_t = [\vec{h}_t, \overleftarrow{h}_t] & (4.6) \end{cases}$$

We pass the final hidden layer of the Bi-LSTM to a multi-layer perceptron (MLP) to obtain the final representation vector ( $V_T$ ) of the paper text.

$$V_T = f_{mlp}([h_t]; \theta_{mlp}) \quad (4.7)$$

where  $f_{mlp}$  denotes a three-layer-MLP, and  $\theta_{mlp}$  denotes the parameters in it.

### Image Feature Extraction

First we extract the figures from each paper using PDFFigures 2.0 [215]. Then we make use of the pre-trained VGG-16 with ImageNet [216] weights to train our *Image Modality Feature Extractor*. The *Image Modality Feature Extractor* consists of an end-to-end 16-layer deep Convolutional Neural Network (CNN) followed by a Multi-Layer Perceptron (MLP-2) network. We freeze the first seven layers of the deep VGG-16 CNN and make the subsequent nine layers trainable. The reason being that the number of images in our articles are not suitable for an end-to-end training, hence we only fine-tune the final layers of the pre-trained VGG-16 with our images extracted from the research articles. The output of the final affine layer of the VGG-16 CNN is the input to MLP-2 which has 3-layers with a final sigmoid layer for classification. Like previous, we use

## 4.5 Multimodal Scope Detection

---

the activations of the preceding fully-connected layer of MLP-2 as the features of Image I. Hence,

$$V_I = f_{mlp}([h_{CNN}]; \theta_{mlp}) \quad (4.8)$$

Where  $V_I$  is the final image representation,  $h_{CNN}$  are the activations of the last hidden layer of VGG-16 CNN,  $f_{mlp}$  denotes a three-layer-MLP, and  $\theta_{mlp}$  denotes the parameters in it. We concatenate all the image representations for a paper to generate the final image representation. If no images are there in a paper, we use a zero-padded vector of dimension equivalent to 8 images as a feature vector for image modality.

We also take the *Bag-of-Words* representation of the **image captions** and fuse it with the corresponding image feature representation via concatenation.

### Why training with VGG-16?

The number of images found in the research papers is not always adequate to train a deep neural feature extractor from scratch. VGG-16 is a deep CNN with 16 layers trained on millions of images. VGG-16 is also a *state-of-the-art* object detector and has been used for transfer learning in many use-cases. Hence we use pre-trained VGG-16 to aid in our high-level feature extraction from the paper images. We freeze the first seven layers as they usually discover low-level features like edges etc.

### Bibliography Feature Extraction

In an earlier work [206] we show that the Bibliography section consists of important domain information regarding the scope of an article to a venue. Especially the citation titles and venues hold significant domain information. Hence we consider *Bibliography* as a separate channel of the text modality here. We find that the vocabulary size of citations for a particular venue (journal/conference) is limited. Hence we proceed with a simple *Bag-of-Words* model to generate the bibliographic feature representations for this channel. To obtain bibliographic feature representation vector  $V_B$  of the document, we concatenate the BoW vectors of bibliographic paper titles ( $t_i$ ) and venues ( $v_i$ ).

$$V_B = BoW(v_i) || BoW(t_i) \quad (4.9)$$

## Phase-II: Attention Weighted Multimodal Classification

### Attention-Based Multimodal Fusion

At this stage we have the feature representations from the three modalities (Full-Text, Image, and Bibliography)<sup>9</sup>. To get the best out of each modality, we make use of *Attention* mechanism [217] popular in deep neural networks. Attention mechanism has the ability to focus on the most important parts of an object relevant to the classification, improving the performance of the baseline deep neural networks. The attention mechanism has been successfully employed in several NLP tasks such as sentiment analysis [218]. The motivation behind using the attention layer is that: *Not all modalities contribute equally to determine the domain of a research article pertaining to a certain venue*. To prioritise only important modalities, we use an attention layer, which takes as an input feature representations from the text, image, and bibliography modalities and outputs an attention score for each modality. Using these scores, the modality contributing more would have higher attention weights. We take the dot product of the respective attention weights with the modality representations and fuse them via concatenation to form the *Attention Weighted Joint Multimodal Feature Representation (AWJMFR)*. The fused multimodal vector  $F$  is computed as follows:

Let  $M_I$ ,  $M_T$ ,  $M_B$  be the feature representation from various modalities where  $M_I=V_I$ ,  $M_T=V_T$ , and  $M_B=V_B$  respectively. The dimensions of  $M_I$ ,  $M_T$ ,  $M_B$  are  $d_I$ ,  $d_T$ ,  $d_B$  respectively.

$$M = [M_I, M_T, M_B] \quad (4.10)$$

$$X = ReLU(W_1^T M) \quad (4.11)$$

$$A = Softmax(W_2^T X) \quad (4.12)$$

Where  $W_1$  and  $W_2$  are the weights of the first and second layer neurons respectively.

Let the attention weights obtained from the three modalities be

$$A = [A_I, A_T, A_B] \quad (4.13)$$

We concatenate the modality vectors after scaling them with attention weights and obtain the final feature fusion AWJMFR:

---

<sup>9</sup>Although a channel of text modality, we consider Bibliography as a separate modality as the text-form in the bibliography is quite different from that in paper body



## 4.5 Multimodal Scope Detection

---

$$F = AWJMFR = [M_I A_I || M_T A_T || M_B A_B] \quad (4.14)$$

where  $||$  signifies concatenation.

### Scope Classification

Finally we pass the AWJMFR through a 3-layer MLP for classification into two classes. We keep *Sigmoid* activation in the final layer as the first task is a binary classification one. For the multi-class problem (Task 2) we keep *Softmax* in the final layer for classification into 3-classes.

#### 4.5.4 Experimental Setup

We discuss the experimental setup in this section.

##### Baselines

To the best of our knowledge, there are no works till date which addresses this problem of multimodal research article classification in the scholarly domain. Hence, we keep the unimodal features (Only Text, Only Image, Only Bibliography) as the baselines for our experiments. Majority of the available journal recommender systems takes *Title* and *Abstract* of a paper as input to suggest a relevant journal. Although our objective is not a recommendation, we also investigate the contributions of individual sections to identify the scope of a candidate paper to a journal.

##### Hyperparameter Details

We enlist the hyperparameter details in accordance to Figure 4.3. The end-to-end trainable Bi-Directional LSTM+MLP in the *Textual Modality Feature Extractor* takes input from the Universal Sentence Encoder. Each sentence has dimension 512 and for each paper we set the number of sentences as 500. The batch size is 64 with binary cross entropy as loss function and *Adam* as the optimizer. The activations in the dense layer is *ReLU* whereas the activation in the final MLP-1 layer is *Sigmoid* for binary classification. We ran 10 epochs until convergence with a learning rate 0.001 and a dropout of 0.3 in MLP-1. Both the Bi-LSTM and MLP-1 has 3 layers. The output of the full-text feature extractor is a representation of 4000 dimension.

The *Image Modality Feature Extractor* comprises of the VGG-16 CNN followed by a 3-layer Multi-layer Perceptron (MLP-2). The input image has dimension  $256 \times 256$ . We freeze the first 7 layers of pre-trained VGG-16, train the remaining nine layers with the input paper images (set to a maximum of 8 per paper). We take the activations of the affine layer as input to MLP-2.

The MLP has *ReLU* activations in dense layer and *Sigmoid* activation in the final layer for classification. The batch size is 128 with binary cross entropy loss and *Adam* optimizer. We continue till ten epochs until convergence with a 0.5 dropout. The output of the *Image Modality Feature Extractor* is a joint representation of images and corresponding captions with dimension  $4096 \times 8 + |d|$  where  $d$  is the length of the image caption vocabulary.

For our Bibliography modality, we follow the simple *Bag-of-Words* model for representing citation title and citation venue. We prune stop words, words with a frequency less than 3 for titles, and less than 6 for venues. The dimension of the output bibliography feature representation depends on the length of the vocabulary for each venue.

For the *Attention* layers in Phase-II of our architecture, we use *ReLU* activations in the dense layer with a dropout of 0.25 and *softmax* in the final layer to learn the attention weights. Further, we use binary cross-entropy as the loss function and *Adam* optimizer with batch size=64 and 20 epochs. The final layer has *Sigmoid* activations for binary classification into *Within Scope* and *Out of Scope*. For *Task 2*, the final layer has *Softmax* activation with categorical cross-entropy as the loss function.

#### 4.5.5 Results and Analysis

We discuss and analyse our results on the two datasets in the subsequent section.

##### Results on Dataset-I

Table 4.4 shows our experimental results on Dataset-I. Here we want to see if our deep neural network can identify *within Scope* and *Out-of-Scope* papers and test it on a dataset comprising six different Computer Science journals. Different modality and section combinations allow us to understand the significance of each modality/section. This also serves as a means of our ablation study.

##### Image Modality

The image modality performs the worst across all the journals. We study the data and find that most of the extracted images are curves/graphs which are generic to all the journals. A major section of those graphical figures is white spaces signifying no object as such. Hence our feature extractor could not discover useful distinguishing features. However, the image+bibliography channel attains a gain of 4%, 3%, 2%, 9%, and 3% over only bibliography in terms of accuracy for JNCA, ARTINT, COMNET, SIMPAT, and CSI respectively. Quite obvious that Statistics and Probability Letters (STATPRO) do not contain enough images and hence image features are not useful here (we observe a significant drop in F-Score values when combined with other

## 4.5 Multimodal Scope Detection

Table 4.4: Scope Detection (Binary Classification) Results on Dataset-I (Elsevier Journals), Accuracy in %

Journals	JNCA		ARTINT		COMNET		SIMPAT		STATPRO		CSI	
	$F_1$	A	$F_1$	A	$F_1$	A	$F_1$	A	$F_1$	A	$F_1$	A
Only Title	0.82	84	0.78	79	0.77	78	0.73	73	0.79	79	0.77	78
Only Abstract	0.82	81	0.87	86	0.89	88	0.79	79	0.88	88	0.84	86
Only Image	0.73	74	0.53	55	0.37	50	0.63	64	0.34	53	0.57	57
Image Captions	0.77	76	0.63	65	0.82	81	0.71	70	0.69	72	0.67	68
Full Text	0.93	89	0.93	93	<b>0.96</b>	<b>95</b>	0.88	88	<b>0.93</b>	<b>93</b>	0.91	93
Bibliography	0.87	86	0.83	86	0.85	84	0.71	72	0.84	85	0.83	83
Image+Abstract	0.85	86	0.89	88	0.88	88	0.81	80	0.82	83	0.85	86
Image+Full-Text	0.93	92	0.93	<b>94</b>	0.95	<b>95</b>	0.88	<b>90</b>	0.85	86	0.92	91
Image+Bibliography	0.92	90	0.89	89	0.86	86	0.79	81	0.85	85	0.85	86
Image+Full-Text+Bib	<b>0.94</b>	<b>95</b>	<b>0.95</b>	<b>94</b>	0.93	<b>95</b>	<b>0.89</b>	<b>90</b>	0.92	<b>93</b>	<b>0.93</b>	<b>94</b>

channels).

However, we argue that the role of images as a differentiator could be more significant for certain biological, natural science, medicine journals where images featuring real-life objects are more pronounced, present in case studies and form a central part of the research.

### Bibliography Channel

We observe that the bibliography channel achieves comparable performance with the *Only Abstract* input. Where sometimes paper abstracts are not sufficient, the bibliography may come to the rescue. Bibliography section holds a good amount of domain information in citation titles and venues which we exploit in our experiments.

### Text Modality

With the sheer volume of information, *Full-Text* is the clear winner, sometimes even better than other modalities combined. When coupled with additional Bibliography and Image information, we observe a gain of 6% (JNCA), 1% (ARTINT), 2% (SIMPAT), 1% (CSI) in terms of accuracy. For COMNET and STATPRO, there is no change. Our attention module emphasised the full-text modality with much higher weights than others. Full-text processing might be computationally expensive, but always there is this trade-off between high accuracy and volume of information processing.

Our best performing model combines all the modalities of information and achieves significant performance improvement over the baselines and individual channels. We observe a gain of more than 10% over individual channels across all journals. Automatically identifying seemingly *out-of-scope* articles is a very crucial yet delicate task. Hence designing a highly accurate system is the need of the problem. Our results clearly suggest that it is required to consider all modalities of information to achieve that goal.

## Results on Dataset-II

Table 4.5 shows our results on the Dataset-II. Here we can observe identical behavior as in Dataset-I, almost resonating with the earlier findings. Although the objective of the task is a bit different than with Dataset-I, we still achieve good performance with our overall model. Basically we try to address, among probable venues, to which venue should a prospective paper go?

### Text Modality

The most contributing modality is again the Full-Text which is quite obvious. Majority of the recent AI, NLP, and CV papers are Machine Learning/Deep Learning based. So most of the technical aspects are very close. For e.g., Convolutional Neural Network (CNN) was widely used for image processing and Computer Vision problems, but recently has shown great success in dealing with NLP problems. Similarly, we can see several other Machine Learning concepts finding foray in NLP and CV papers. Still the scope of those papers could be differentiated with the problem they address, the data they work on, and the insights they derive.

### Bibliography Channel

However, we see that Bibliography channel fares almost close to the Full-Text modality. This is because the type of citations for NLP and CV would be different. AI/ML papers are based on core mathematical and theoretical groundings, many are from disciplines other than NLP, CV; hence have a different category of bibliographic citations in comparison to more application oriented NLP and CV papers.

### Image Modality

Image modality features alone do not perform well. But when augmented with the paper abstract gains an accuracy of more than 14% (at least). Even combination of Bibliography channel with Image features reaches a competitive benchmark as Full-Text. The less performance of images is because images present in papers are not uniform in terms of numbers, quality, etc. Many of them are graphs which convey little information about the domain of the manuscript, as we discuss earlier too.

### Error Analysis

Although few, errors in our system are due to:

1. We were not able to process the text crisply as is there in the paper. Parsing errors, lot of

## 4.5 Multimodal Scope Detection

---

Table 4.5: Results on Dataset-II (AI/ML/NLP/CV). Multi-class classification.

Journals	AI/ML		CV		NLP	
	$F_1$	Acc	$F_1$	Acc	$F_1$	Acc
<b>Only Title</b>	0.75	74%	0.79	80%	0.85	84%
<b>Only Abstract</b>	0.76	71%	0.83	84%	0.87	90%
<b>Only Image</b>	0.75	70%	0.62	67%	0.79	75%
<b>Image Captions</b>	0.65	52%	0.75	78%	0.68	65%
<b>Full Text</b>	0.92	93%	0.92	91%	0.93	93%
<b>Bibliography</b>	0.87	85%	0.90	91%	0.92	94%
<b>Img+Abs</b>	0.95	<b>95%</b>	0.91	92%	0.92	92%
<b>Img+FT</b>	<b>0.96</b>	<b>95%</b>	0.93	92%	0.93	<b>96%</b>
<b>Img+Bib</b>	0.86	83%	0.88	92%	<b>0.94</b>	95%
<b>Img+FT+Bib</b>	<b>0.96</b>	<b>95%</b>	<b>0.94</b>	<b>93%</b>	<b>0.94</b>	93%

*out-of-vocabulary* words are few reasons. We should have used embeddings generated from scholarly data.

2. Majority of the images (graphs) were similar for all the classes. Less amount of distinctive images.
3. Overlapping nature of textual content (in case of Dataset-II). For e.g., similar technologies used in NLP, CV papers. At least considering the surface form of the texts.

### 4.5.6 Conclusions

Here in this work, we conduct a thorough study of the role of different modalities and information channels for determining the belongingness of an article to a venue. To the best of our knowledge, we are the first to employ a deep neural network for the problem in hand. Our extensive experiments on an array of journals and conferences show that a highly accurate system of this kind is possible. The definition of scope for each venue is different and is not always dependant on specific topics of interest. Our network learns the scope characteristics of each venue and corresponding *domain of operation* from past accepted papers. With our experiments on the Dataset-II, we are able to address the place holder of a manuscript in highly related venues. This research could be suitably moulded to build a venue recommender system for the authors as well. For the editors, it would be much easier to identify potential misfit submissions and intimate the authors quickly thus accelerating the overall peer review system.

## 4.6 Multiview Clustering for Scope Detection of Scientific Articles

Desk-Rejection is an unfortunate yet common occurrence in academic peer review. In spite of having merit, many papers suffer rejections since they do not fall within the scope of the intended journal. Studies [67, 206] show that around 25-30% of desk-rejections owe to misfit submissions. Editors invest a considerable amount of time in judging the appropriateness of submissions at the desk before forwarding it to reviewers for meticulous evaluation. Here in this work we investigate if we can isolate potential *out-of-scope* submissions from clusters of *in-scope* papers. Our idea is simple: articles which are within the scope of a journal would be similar in some aspects, share common keywords, bibliography and hence could be grouped into clusters. Articles which are supposedly *out-of-scope* to that journal would be distant from the clusters of those *in-scope* articles. We adopt a semi-supervised approach to this problem. A journal may have several topics of interest. In the first phase, we use a portion of the past accepted (labelled *in-scope* data) papers to create the various clusters representing topically similar papers. In the second phase, we take a set of unlabelled data points (research papers) and further cluster them into two groups: *In-Scope* and *Out-of-Scope*. The clusters in the first phase supervise the clustering in the second phase. To understand the domain of a research article we view it from three different perspectives: *lexical*, *semantic*, and *bibliography*. Earlier we show that along with the full-text information, the bibliography section consists of important information (citation titles, venues) regarding the domain of a paper [208]. The contributions of the current work are:

- *We present an effective approach that takes advantage of multiple views of a research article which together describe the appropriateness of the article to a journal.*
- *Our semi-supervised approach learns from a less amount of training data yet yields high performance. It can prove very effective for new journals where past data is less or if the field is rapidly evolving.*

With our current approach based on semi-supervised multiview-clustering, we show that we achieve comparable performance with the deep architectures with much less data to bootstrap the process. We differ from our earlier approaches as we consider the *low-resource* and *cold start* scenario here. Our treatment that an *outlier* paper to the cluster of accepted articles could be *out-of-scope* proves helpful as evident from our results we discuss in our subsequent sections.

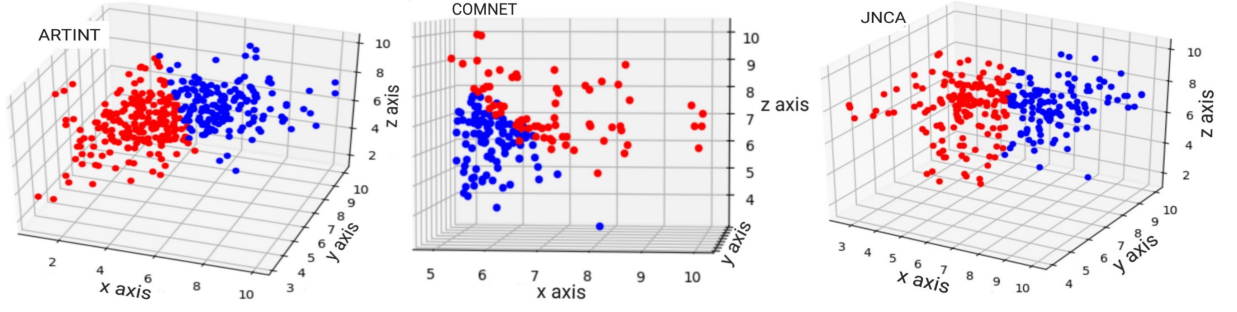


Figure 4.4: Multi-view Clustering of *In-Scope* and *Out-of-Scope* articles, X-axis→Semantic View, Y-axis→Lexical View, Z-axis→Bibliography View

### 4.6.1 Dataset Description/Preprocessing

We take data from three Elsevier Computer Science journals: Artificial Intelligence (ARTINT), Computer Networks (COMNET), and Journal of Network and Computer Applications (JNCA). For the first phase, we use 200 recently published articles (*In-Scope*) from each journal. For the second phase, we use 244, 175, 169 *In-Scope* articles and 154, 160, 126 *Out-of-Scope* articles from ARTINT, COMNET, and JNCA, respectively. We convert the articles from PDF to .json using the Science Parse<sup>10</sup> library for information extraction from full-text and bibliography.

### 4.6.2 Methodology

We represent the paper with the following views: *Lexical* view corresponds to the surface form of the text; *Semantic* view would delve into the meaning representation of the full-text; *Bibliographic* view would take into account the type of citations the paper contain. We use feature representations from all of these three views and remove stop words/irrelevant words from the scholarly texts.

- **Semantic View:** We select top 30 frequent words from each article (A). We extract the keywords from the article using the RAKE keyword extraction algorithm (B). Finally, we perform  $T = A \cup B$ . To get the semantic representation of T, we take the *word2vec* [163] representations of individual words present in T and concatenate them to form the semantic document representation. We generate *word2vec* word vectors from 400k Elsevier Computer Science journal papers [206] to preserve scholarly domain knowledge.
- **Lexical View:** We adopt similar approach and use *term frequency-inverse document frequency* (*tf-idf*) as the lexical representation.

<sup>10</sup><https://github.com/allenai/science-parse>

- **Bibliography View:** We use only the *Citation Title* and *Citation Venue* and use *tf-idf* to generate the bibliography representation.

Next, we use the multi-view Archived MultiObjective Simulated Annealing (AMOSa) clustering algorithm with default parameters [219] considering these three views to generate the consensus document clusters<sup>11</sup>. Finally we make use of K-Medoids algorithm upon the consensus clusters to separate the input data into *In-Scope* and *Out-of-Scope* groups. Algorithm 2 details the procedure.

---

**Algorithm 2** Multiview Clustering

---

**INITIALIZE**

- 1: Given an input dataset of N *In-Scope* Papers
- 2: Extract feature representations corresponding to the three views (Semantic, Lexical, Bibliography)

**PHASE-I**

- 3: Execute AMOSA algorithm to generate the consensus partitioning satisfying multiple views. The dataset is partitioned into k consensus clusters ( $U_1, U_2, \dots, U_k$ )
- 4: Using the three different representations, find the center (mean) ( $U_{ci}$ ) of each consensus cluster  $U_i$  (i=1 to k)

**PHASE-II**

- 5: Given m unlabelled samples ( $X_1, X_2, \dots, X_m$ ), where  $X \in [G1 \cup G2]$ ;  $G1: In-scope$ ,  $G2: Out-of-Scope$
  - 6: For each  $X_i$ , find the closest cluster  $U_j$  using Mahalanobish distance. Let  $dS_i, dL_i, dB_i$  be the distances between  $X_i$  and  $U_{cj}$  for the three different views.
  - 7: Represent each  $X_i$  by three distance features  $dS_i, dL_i, dB_i$
  - 8: Normalize ( $dS_i, dL_i, dB_i$ ) using Min/Max normalization
  - 9: Apply K-Medoids to find two partitions ( $X_{C1}, X_{C2}$ ) of unlabelled data X
  - 10: Find the centers ( $C_{C1}, C_{C2}$ ) of the two clusters  $X_{C1}$  and  $X_{C2}$
  - 11: Find the Euclidean distances ( $d1, d2$ ) from origin to  $C_{C1}$  and  $C_{C2}$  respectively
  - 12: If  $d1 < d2$ :  $X_{C1}$  is the *In-Scope*,  $X_{C2}$  is the *Out-of-Scope* cluster
  - 13: Else:  $X_{C2}$  is the *In-Scope* cluster and  $X_{C1}$  is the *Out-of-Scope* cluster
- 

### 4.6.3 Evaluation

In Phase-I we use labelled data elements (Only In-Scope) to build up our consensus clusters. In Phase-II we input unlabelled data elements; however, we know the ground-truth (Semi-Supervised). We evaluate our approach by observing the accuracy of the predicted cluster labels of the unlabelled input data points. The objective is simple: *How many unlabelled input elements are correctly put into corresponding groups?* This gives the *belongingness* of a data point to the certain cluster. Figure 4.4 shows the efficiency of our approach tested on the three different

---

<sup>11</sup>Consensus clustering, also called aggregation of clustering (or partitions), refers to the situation in which a number of different (input) clusterings have been obtained for a particular dataset and it is desired to find a single (consensus) clustering which is a better fit in some sense than the existing clusterings.



## 4.7 Insights gained from this work

---

Table 4.6: Cluster Prediction (*In-Scope* or *Out-Scope*) Results on the 3 journals,  $\dagger \rightarrow$  *Baselines*

Journals $\rightarrow$	ARTINT		COMNET		JNCA	
	$F_1$	Acc	$F_1$	Acc	$F_1$	Acc
<b>Lexical<math>\dagger</math></b>	0.66	40.20	0.56	55.19	0.60	44.39
<b>Semantic<math>\dagger</math></b>	0.72	43.21	0.65	54.62	0.30	42.03
<b>Bibliography<math>\dagger</math></b>	0.71	61.05	0.63	56.71	0.65	51.86
<b>Lex+Sem</b>	0.72	62.31	0.71	62.68	0.79	74.57
<b>Lex+Bib</b>	0.64	48.49	0.58	45.37	0.66	53.89
<b>Sem+Bib</b>	0.70	58.29	0.72	65.37	0.76	69.49
<b>Sem+Lex+Bib</b>	<b>0.95</b>	<b>94.91</b>	<b>0.97</b>	<b>97.61</b>	<b>0.94</b>	<b>93.22</b>

journals. We are able to form distinct clusters for *In-Scope* and *Out-Scope* data points after Phase-II. Table 4.6 shows our results for the predicted cluster labels against the actual labels. We find that the *Bibliography* view is the most effective one. However, the multiview approach yields high performance justifying our assumption that the three views are important to identify the belongingness of the article to the scope of the journal.

### 4.6.4 Conclusion

Here we explore multiview clustering to identify the appropriateness of an article to a journal. We see that, with little supervision, our method shows promise to address scope detection of research articles leveraging on multiple views of the article concerned.

## 4.7 Insights gained from this work

As is evident from our feature engineering approach, we found the *bibliography* section to hold a good amount of domain information. Our multi-modal deep neural architecture reveals that using full paper text we can achieve the best performance for classifying research articles according to their domain. However, full-paper texts are not always available due to variety of reasons. Paper meta-data including embedded images could help in such cases. Supposedly, role of images would be more prominent in cases of journals from Medicine, Biology, Astrophysics, etc. where images within papers play a pivotal role. The third approach based on multiview clustering shows that exploiting the multiple views of a research paper we can achieve comparable performance with respect to data hungry deep methods. It is comparably easier to imagine a *out-of-scope* paper as an outlier to the cluster of accepted articles.

## 4.8 Limitations

We could see that our scope detection methods are highly accurate on a limited dataset of scientific papers. The major limitations are processing of full-text papers and associated artifacts

as part of the approaches. The feature engineering based solution might be a better fit here than the deep architectures. Another limitation of our approaches is that these would not work correctly for interdisciplinary venues which accepts papers from multiple domains. Also, if a paper is within domain but is consisting novel content which is not that much evident in earlier issues of the venue, our approach may fail to classify it correctly. A novel paper may get penalized with our approach and can be treated as an outlier.

## 4.9 Chapter Summary

In this chapter we address an important problem related to peer review in scholarly communications. Automatic *Scope Detection* of research articles would help both the authors and the editorial team to reduce the turn-around time in peer review. Our approaches are simple consisting of simple feature based methods to multiview clustering to deep neural architectures. We re-emphasize that we do not support a fully automated system in peer review, but AI as an assistant in the process to help editors in their decisions.

# AI in Peer Review and Finding a Research Lineage: Tackling the Scientific Burden of Information Explosion

---

○

In this chapter, we investigate three different problems pertaining to peer review and literature-based knowledge discovery. The first problem deals with predicting the outcome of a peer review process on the basis of paper full-text and human-written reviews. The second problem follows from the first and we investigate if the peer review was significant enough for decision-making. With the third problem we explore how significant was a piece of research in the community via tracing its significant citations. The following sections details about each problem and our approach towards those.

---

○

## 5.1 Introduction

This chapter is a culmination of our attempt towards certain problems for having AI assistance in the peer review process. There has already been several attempts to reduce the information overload in scholarly communications. The deluge of scientific information from the thousands of papers getting added to the scholarly knowledge base adds complexity to the search and retrieval of relevant and high-quality information. Two of the most important areas under impact of this scientific information explosion are: the *peer review process* and *relevant literature discovery*. In this chapter we report our investigations on what if we train an AI to predict decisions based on past accepted/rejected papers along with the corresponding reviewer comments? *Can AI act as the fourth reviewer here?* How do we leverage on the sentiment of the reviewer to know the in-general consensus of the reviewers towards the article under scrutiny? While we investigated the impact of AI in decision-making, we realized that before automatically predicting the decision based on the reviews, it makes more sense to have trust on the evaluation, i.e. the peer reviews. Only if the peer reviews are reliable enough to reflect the correct evaluation of the paper, we can attempt to automate the decision-making. The area-chair/program-chair in a conference or the editor of a journal plays this crucial role. Hence, investigating the automatic evaluation of peer review quality became one direction of our research organically.

To reduce the scholarly information overload on the reviewers (researchers) to review a paper, we probed into how we can automate relevant literature discovery. We leveraged on the idea of information propagation in a citation network via identifying *meaningful citations*. Hence, this chapter is dedicated towards investigating problems that would help reduce the scholarly information overload in the peer review process.

## 5.2 Predicting Peer Review Outcome

Automatically validating a research artefact is one of the frontiers in Artificial Intelligence (AI) that directly brings it close to competing with human intellect and intuition. Although criticized sometimes, the existing peer review system still stands as the benchmark of research validation. The present-day peer review process is not straightforward and demands profound domain knowledge, expertise, and intelligence of human reviewer(s), which is somewhat elusive with the current state of AI. However, the peer review texts, which contains rich sentiment information of the reviewer, reflecting his/her overall attitude towards the research in the paper, could be a valuable entity to predict the acceptance or rejection of the manuscript under consideration. Here in this work, we investigate the role of reviewers sentiments embedded within

## 5.2 Predicting Peer Review Outcome

---

peer review texts to predict the peer review outcome.

The rapid increase in research article submissions across different venues is posing a significant management challenge for the journal editors and conference program chairs<sup>1</sup>. Among the load of works like assigning reviewers, ensuring timely receipt of reviews, slot-filling against the non-responding reviewer, taking informed decisions, communicating to the authors, etc., editors/program chairs are usually overwhelmed with many such demanding yet crucial tasks. However, the major hurdle lies in to decide the acceptance and rejection of the manuscripts based on the reviews received from the reviewers.

The quality, randomness, bias, inconsistencies in peer reviews is well-debated across the academic community [220]. Due to the rise in article submissions and non-availability of expert reviewers, editors/program chairs are sometimes left with no other options than to assign papers to the novice, out of domain reviewers which sometimes results in more inconsistencies and poor quality reviews. To study the arbitrariness inherent in the existing peer review system, organisers of the NIPS 2014 conference assigned 10% submissions to two different sets of reviewers and observed that the two committees disagreed for more than quarter of the papers [221]. Again it is quite common that a paper rejected in one venue gets the cut in another with little or almost no improvement in quality. Many are of the opinion that the existing peer review system is fragile as it only depends on the view of a selected few [222]. Moreover, even a preliminary study into the inners of the peer review system is itself very difficult because of data confidentiality and copyright issues of the publishers. However, the silver lining is that the peer review system is evolving with the likes of OpenReviews<sup>2</sup>, author response periods/rebuttals, increased effective communications between authors and reviewers, open access initiatives, peer review workshops, review forms with objective questionnaires, etc. gaining momentum.

The PeerRead dataset [5] is an excellent resource towards research and study on this very impactful and crucial problem. With our ongoing effort towards the development of an Artificial Intelligence (AI)-assisted peer review system, we are intrigued with: *What if there is an additional AI reviewer which predicts decisions by learning the high-level interplay between the review texts and the papers? How would the sentiment embedded within the review texts empower such decision-making?* Although editors/program chairs usually go by the majority of the reviewer recommendations, they still need to go through all the review texts corresponding to all the submissions. A good use case of this research would be: slot-filling the missing reviewer, providing an additional perspective to the editor in cases of contrasting/borderline reviews. This work in no way attempts to replace the human reviewers; instead, we are intrigued to see how

---

<sup>1</sup>Apparently CVPR, NIPS, AAAI 2019 received over 5100, 4900, 7000 submissions respectively!

<sup>2</sup><https://openreview.net>

an AI can act as an additional reviewer with inputs from her human counterparts and aid the decision-making in the peer review process.

### 5.2.1 Problem Definition

Simply stating, *given a paper  $P$  and reviews  $R1, R2, R3$ , can we automatically predict the recommendation-score of  $R1, R2, R3$  for  $P$ ? Can we also further predict the final-decision regarding  $P$ ?*

We develop a deep neural architecture incorporating full paper information and review text along with the associated sentiment to predict the acceptability and recommendation score of a given research article. Our proposed deep neural architecture takes into account three channels of information: the paper, the corresponding reviews, and the review polarity to predict the overall recommendation score as well as the final decision. We achieve significant performance improvement over the baselines ( $\sim 29\%$  error reduction) proposed in PeerRead. An AI of this kind could assist the editors/program chairs as an additional layer of confidence in the final decision making, especially when non-responding/missing reviewers are frequent in present day peer review. We perform two tasks, a classification (predicting accept/reject decision) and a regression (predicting recommendation score) one. The evaluation shows that our proposed model successfully outperforms the earlier reported results in PeerRead. We also show that the addition of review sentiment component significantly enhances the predictive capability of such a system.

### 5.2.2 Data Description and Analysis

The PeerRead dataset consists of papers, a set of associated peer reviews, and corresponding accept/reject decisions with aspect specific scores of papers collected from several top-tier Artificial Intelligence (AI), Natural Language Processing (NLP) and Machine Learning (ML) conferences. Table 5.1 shows the data we consider in our experiments. We could not consider NIPS and arXiv portions of PeerRead due to the lack of aspect scores and reviews, respectively. For more details on the dataset creation and the task, we request the readers to refer to [5]. We further use the submissions of ICLR 2018, corresponding reviews and aspect scores to boost our training set for the decision prediction task. One motivation of our work stems from the finding that aspect scores for certain factors like *Impact*, *Originality*, *Soundness/Correctness* which are seemingly central to the merit of the paper, often have very low correlation with the final recommendation made by the reviewers as is made evident in [5]. However, from the heatmap in Figure 5.1 we can see that the reviewer’s sentiments (compound/positive) embedded within the review texts

## 5.2 Predicting Peer Review Outcome

have visible correlations with the aspects like *Recommendation*, *Appropriateness* and *Overall Decision*. This also seconds our finding that determining the scope or appropriateness of an article to a venue is the first essential step in peer review [208]. Since our study aims at deciding the fate of the paper, we take predicting recommendation score and overall decision as the objectives of our investigation. Thus our proposal to augment sentiment of reviews to the deep neural architecture seems intuitive.

Table 5.1: Dataset Statistics

Venues	#Papers	#Reviews	Aspect	Acc/Rej
ICLR 2017	427	7270	Y	172/255
ACL 2017	137	275	Y	88/49
CoNLL 2016	22	39	Y	11/11
ICLR 2018	909	2741	Only Rec	336/573
Total	1495	10325	—	607/888

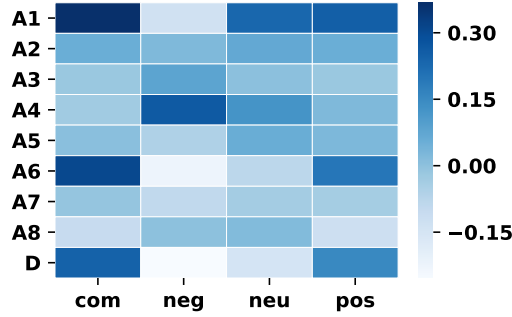


Figure 5.1: Pearson Correlation of Review Sentiment (:X) with different Aspect Scores (:Y) on ACL 2017 dataset. A1→Appropriateness, A2→Clarity, A3→Impact, A4→Meaningful Comparison, A5→Originality, A6→Recommendation, A7→Soundness/Correctness, A8→Substance, D→Decision. pos→Positive Sentiment Score, neg→Negative Sentiment Score, neu→Neutral Sentiment Score, com→Compound Sentiment Score. To calculate the sentiment polarity of a review text, we take the average of the sentence wise sentiment scores from Valence Aware Dictionary and sEntiment Reasoner (VADER) [7].

### 5.2.3 Methodology

#### Pre-processing

At the very beginning, we convert the papers in PDF to .json encoded files using the Science Parse<sup>3</sup> library.

<sup>3</sup><https://github.com/allenai/science-parse>

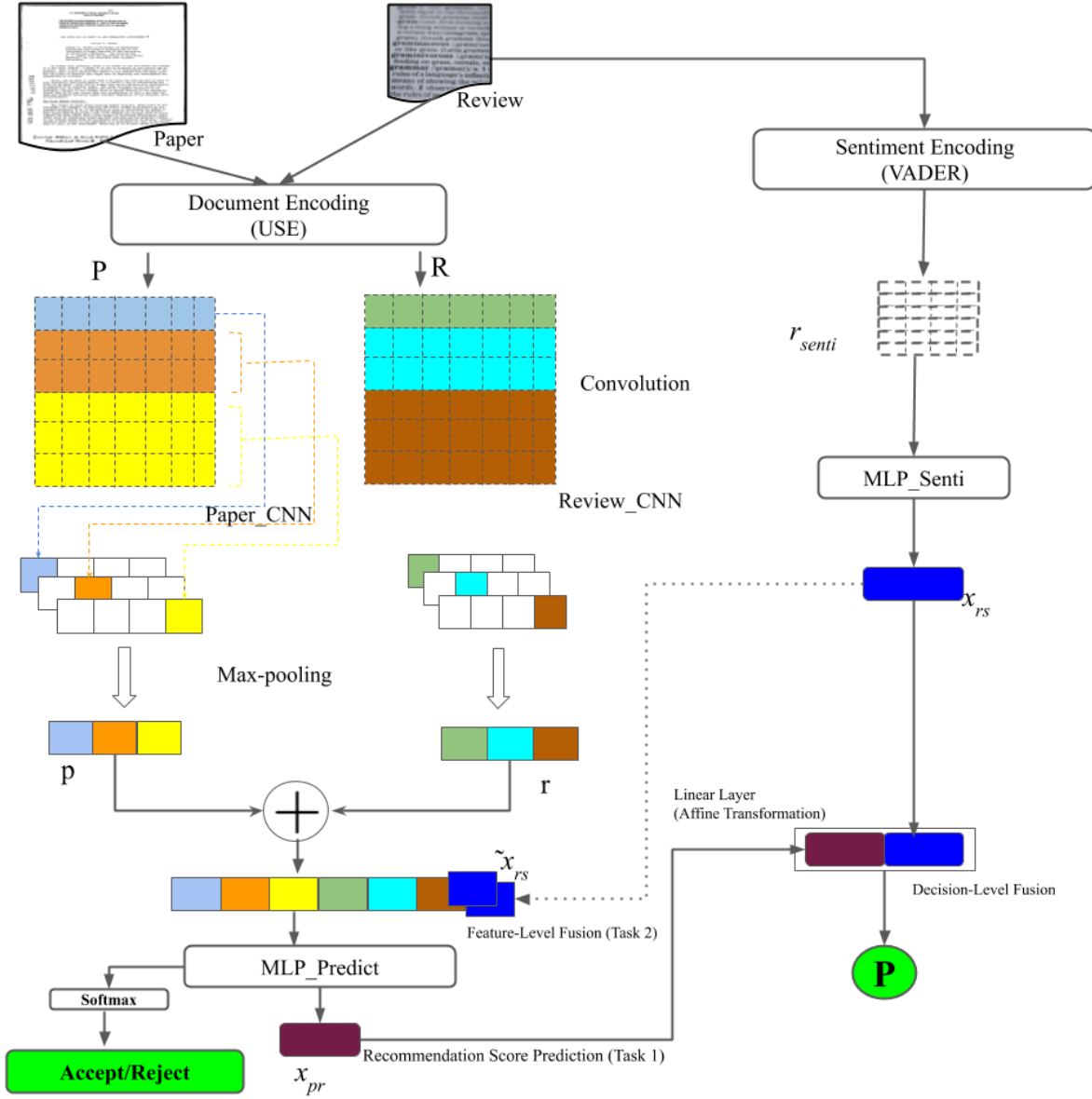


Figure 5.2: *DeepSentiPeer*: A Sentiment Aware Deep Neural Architecture to Predict Reviewer Recommendation Score. Decision-Level Fusion and Feature-Level Fusion of Sentiment are shown for Task 1 and Task 2, respectively.

### *DeepSentiPeer* Architecture

Figure 5.2 illustrates the overall architecture we employ in our investigation. The left segment is for the decision prediction while the right segment predicts the overall recommendation score.

### Document Encoding

We extract full-text sentences from each research article and represent each sentence  $s_i \in \mathbb{R}^d$  using the Transformer variant of the Universal Sentence Encoder (USE) [198],  $d$  is the dimension



## 5.2 Predicting Peer Review Outcome

---

of the sentence semantic vector which is 512. A paper is then represented as,

$$\mathbf{P} = \mathbf{s}_1 \oplus \mathbf{s}_2 \oplus \dots \oplus \mathbf{s}_{n_1}, \mathbf{P} \in \mathbb{R}^{n_1 \times d} \quad (5.1)$$

$\oplus$  being the concatenation operator,  $n_1$  is the maximum number of sentences in a paper text in the entire dataset (padding is done wherever necessary). Similarly, we do this for each of the reviews and create a review representation as

$$\mathbf{R} = \mathbf{s}_1 \oplus \mathbf{s}_2 \oplus \dots \oplus \mathbf{s}_{n_2}, \mathbf{R} \in \mathbb{R}^{n_2 \times d} \quad (5.2)$$

$n_2$  being the maximum number of sentences in the reviews.

### Sentiment Encoding

The sentiment encoding of the review is done using VADER Sentiment Analyzer. For a sentence  $s_i$ , VADER gives a vector  $\mathbf{S}_i$ ,  $\mathbf{S}_i \in \mathbb{R}^4$ . The review is then encoded (padded where necessary) for sentiment as

$$\mathbf{r}_{senti} = \mathbf{S}_1 \oplus \mathbf{S}_2 \oplus \dots \oplus \mathbf{S}_{n_2}, \mathbf{r}_{senti} \in \mathbb{R}^{n_2 \times 4}. \quad (5.3)$$

### Feature Extraction with Convolutional Neural Network

We make use of a Convolutional Neural Network (CNN) to extract features from both the paper and review representations. The convolution operation works by sliding a filter  $\mathbf{W}_{f_k} \in \mathbb{R}^{l \times d}$  to a window of length  $l$ , the output of such  $h^{th}$  window is given as,

$$\mathbf{f}_h^k = g(\mathbf{W}_{f_k} \cdot \mathbf{X}_{h-l+1:h} + b_k) \quad (5.4)$$

$\mathbf{X}_{h-l+1:h}$  means the  $l$  sentences within the  $h^{th}$  window in Paper  $\mathbf{P}$ .  $b_k$  is the bias for the  $k^{th}$  filter,  $g()$  is the non-linear function. The feature map  $\mathbf{f}^k$  for the  $k^{th}$  filter is then obtained by applying this filter to each possible window of sentences in the  $\mathbf{P}$  as

$$\mathbf{f}^k = [\mathbf{f}_1^k, \mathbf{f}_2^k, \dots, \mathbf{f}_h^k, \dots, \mathbf{f}_{n_1-l+1}^k], \mathbf{f}^k \in \mathbb{R}^{n_1-l+1}. \quad (5.5)$$

We then apply a max-pooling operation to this filter map to get the most significant feature,  $\hat{\mathbf{f}}^k$  as  $\hat{\mathbf{f}}^k = \max(\mathbf{f}^k)$ . For a paper  $\mathbf{P}$ , the final output of this convolution filter is then given as

$$\mathbf{p} = [\hat{\mathbf{f}}^1, \hat{\mathbf{f}}^2, \dots, \hat{\mathbf{f}}^k, \dots, \hat{\mathbf{f}}^F], \mathbf{p} \in \mathbb{R}^F, \quad (5.6)$$

$F$  is the total number of filters used. In the same way, we can get  $\mathbf{r}$  as the output of the convolution operator for the Review  $\mathbf{R}$ .

We call the outputs  $\mathbf{p}$  and  $\mathbf{r}$  as the high-level representation feature vector of the paper and the review, respectively. We then concatenate these feature vectors (Feature-Level Fusion). The reason we extract features from both is to simulate the editorial workflow, wherein ideally, the editor/chair would look at both into the paper and the corresponding reviews to arrive at a judgement.

### Multi-layer Perceptron

We employ a Multi-Layer Perceptron ( $MLP\_Predict$ ) to take the joint paper+review representations  $\mathbf{x}_{pr}$  as input to get the final representation as

$$\mathbf{x}_{pr} = f_{MLP\_Predict}(\theta_{predict}; [\mathbf{p}, \mathbf{r}]), \quad (5.7)$$

where  $\theta_{predict}$  represents the parameters of the  $MLP\_Predict$ . We also extract features from the review sentiment representation  $\mathbf{x}_{rs}$  via another MLP ( $MLP\_Senti$ ).

$$\mathbf{x}_{rs} = f_{MLP\_Senti}(\theta_{senti}; \mathbf{r}_{senti}), \quad (5.8)$$

$\theta_{senti}$  being the parameters of  $MLP\_Senti$ . Finally, we fuse the extracted review sentiment feature and joint paper+review representation together to generate the overall recommendation score (Decision-Level Fusion) using the affine transformation as

$$prediction = (\mathbf{W}_d \cdot [\mathbf{x}_{pr}, \mathbf{x}_{rs}] + b_d). \quad (5.9)$$

We minimize the Mean Square Error (MSE) between the actual and predicted recommendation score. The motivation here is to augment the human judgement (review+embedded sentiment) regarding the quality of a paper in decision making. The long-term objective is to have the AI learn the notion of good and bad papers from the human perception reflected in peer reviews in correspondence with paper full-text.

### Accept/Reject Decisions

Instead of training the deep network on overall recommendation scores, we train the network with the final decisions of the papers in a classification setting. The entire setup is same but we concatenate all the reviews of a particular paper together to get the review representation. And rather than doing decision-level fusion, we perform feature-level fusion where the decision

## 5.2 Predicting Peer Review Outcome

---

is given as

$$\mathbf{x}_{prs} = f_{MLP\_Predict}(\theta; [\mathbf{p}, \tilde{\mathbf{r}}, \tilde{\mathbf{x}}_{rs}]) \quad (5.10)$$

$$\mathbf{c} = Softmax(\mathbf{W}_c \cdot \mathbf{x}_{prs} + b_c), \quad (5.11)$$

where  $\mathbf{c}$  is the output classification distribution across accept or reject classes.  $\tilde{\mathbf{r}}$  is the high-level representation of review text after concatenating all reviews corresponding to a paper and  $\tilde{\mathbf{x}}_{rs}$  is the output of *MLP\_Senti* on the concatenated review text. We minimize Cross-Entropy Loss between predicted  $\mathbf{c}$  and actual decisions.

### 5.2.4 Experimental Setup

As we mention earlier, we undertake two tasks:

**Task 1:** *Predicting the overall recommendation score* (Regression) and

**Task 2:** *Predicting the Accept/Reject Decision* (Classification).

To compare with [5], we keep the experimental setup (train vs test ratio) identical and re-implement their codes to generate the comparing figures. However, [5] performed Task 2 on ICLR 2017 dataset with handcrafted features, and Task 1 in a deep learning setting. Since our approach is a deep neural network based, we crawl additional paper+reviews from ICLR 2018 to boost the training set.

For Task 1,  $n_1$  is 666 and  $n_2$  is 98 while for Task 2,  $n_1$  is 1494 and  $n_2$  is 525. We employ a grid search for hyperparameter optimization. For Task 1,  $F$  is 256,  $l$  is 5. ReLU is the non-linear function  $g()$ , learning rate is 0.007. We train the model with SGD optimizer, set momentum as 0.9 and batch size as 32. We keep dropout at 0.5. We use the same number of filters with the same kernel size for both paper and review. In Task 2, for Paper.CNN  $F$  is 128,  $l$  is 7 and for Review.CNN  $F$  is 64 and  $l$  is 5. Again we train the model with Adam Optimizer, keep the batch size as 64 and use 0.7 as the dropout rate to prevent overfitting. We intentionally keep our CNN/MLP shallow due to less training data.

### 5.2.5 Results and Analysis

Table 5.2 and Table 5.3 show our results for both the tasks. We propose a simple but effective architecture in this work since our primary intent is to establish that a sentiment-aware deep architecture would better suit these two problems. For **Task 1**, we can see that our review sentiment augmented approach outperforms the baselines and the comparing systems by a wide margin ( $\sim 29\%$  reduction in error) on the ICLR 2017 dataset. With only using review+sentiment information, we are still able to outperform [5] by a margin of 11% in terms of RMSE. A

Table 5.2: Results on Aspect Score Prediction Task. Training is done with only ICLR 2017 papers/reviews,  $\dagger \rightarrow$  Cross-Domain: Training on ICLR and testing upon entire data of ACL/-CoNLL available in PeerRead dataset,  $\ddagger \rightarrow$  Test set is kept the same as [5], RMSE $\rightarrow$ Root Mean Squared Error. CNN variant as in [5] is used as the comparing system.

Baselines	Task 1 $\rightarrow$	Aspect Score Prediction (RMSE)		
	Test Datasets $\rightarrow$	ICLR $\ddagger$	ACL $\dagger$	CoNLL $\dagger$
	Approaches $\downarrow$	2017	2017	2016
Comparing Systems	Majority Baseline	1.6940	2.7968	2.9133
	Mean Baseline	1.6095	2.4900	2.6086
	Only Paper [5]	1.6462	2.7278	3.0591
	Only Review [5]	1.6955	2.7062	2.7072
	Paper+Review [5]	1.6496	2.5011	2.9734
Proposed Architecture <i>DeepSentiPeer</i>	Only Review	1.5812	2.7191	2.6537
	Review+Sentiment	<b>1.4521</b>	2.6845	<b>2.5524</b>
	Paper+Review+Sentiment	<b>1.1679</b>	<b>2.3790</b>	<b>2.5399</b>

Table 5.3: Results on Accept/Reject Classification Tasks. Training is done with ICLR 2017+ICLR 2018 papers/reviews,  $\dagger \rightarrow$  Cross-Domain: Training on ICLR and testing upon the entire data of ACL/CoNLL,  $\ddagger$ Test Set is kept the same as [5], RMSE $\rightarrow$ Root Mean Squared Error,  $*$   $\rightarrow$ 65.79% if only trained with ICLR 2017, Comparing System [5] is feature-based and considers only paper, and not the reviews.

Baseline	Task 2 $\rightarrow$	Accept/Reject (Accuracy)		
	Test Datasets $\rightarrow$	ICLR $\ddagger$	ACL $\dagger$	CoNLL $\dagger$
	Approaches $\downarrow$	2017	2017	2016
Comparing System	Majority Baseline	60.52	33.33	39.94
	Only Paper [5]	55.26*	35.93	41.23
Proposed Architecture <i>DeepSentiPeer</i>	Only Review	65.35	57.12	62.91
	Review+Sentiment	<b>69.79</b>	<b>59.31</b>	<b>62.22</b>
	Paper+Review+Sentiment	<b>71.05</b>	<b>64.76</b>	<b>67.71</b>

further relative error reduction of 19% with the addition of paper features strongly suggests that only review is not sufficient for the final recommendation. A joint model of the paper content and review text (the human touch) augmented with the underlying sentiment would efficiently guide the prediction. For **Task 2**, we observe that the handcrafted feature-based system by [5] performs inferior compared to the baselines. This is because the features were very naive and did not address the complexity involved in such a task. We perform better with a relative improvement of 28% in terms of accuracy, and also our system is end-to-end trained. Presumably, to some extent, our deep neural network learned to distinguish between the probable accept versus probable reject by extracting useful information from the paper and review texts.

## Cross-Domain Experiments

With the additional (but less) data of ACL 2017 and CoNLL 2016 in PeerRead, we perform the cross-domain experiments. We do training with the ICLR data (core Machine Learning papers) and take the test set from the NLP conferences (ACL/CoNLL). NLP nowadays is mostly machine learning (ML) centric, where we find several applications and extensive usage of ML algorithms

## 5.2 Predicting Peer Review Outcome

---

to address different NLP problems. Here we observe a relative error reduction of 4.8% and 14.5% over the comparing system for ACL 2017 and CoNLL 2016, respectively (Table 5.2). For the decision prediction task, the comparing system performs even worse, and we outperform them by a considerable margin of 28% (ACL 2017) and 26% (CoNLL 2017), respectively (Table 5.3). The reason is that the work reported in [5] relies on elementary handcrafted features extracted only from the paper; does not consider the review features whereas we include the review features along with the sentiment information in our deep neural architecture. However, we also find that our approach with only Review+Sentiment performs inferior to the Paper+Review method in [5] for ACL 2017. This again seconds that inclusion of paper is vital in recommendation decisions. Only paper is enough for a human reviewer, but with the current state of AI, an AI reviewer would need the supervision of her human counterparts to arrive at a recommendation. So our system is suited to cases where the editor needs an additional judgment regarding a submission (such as dealing with missing/non-responding reviewers, an added layer of confidence with an AI which is aware of the past acceptances/rejections of a specific venue).

### Analysis: Effect of Sentiment on Reviewer’s Recommendation

Figure 5.3 shows the output activations<sup>4</sup> from the final layer of *MLP\_Senti* against the predicted recommendation scores. We can see that the papers are discriminated into visible clusters according to their recommendation scores. This proves that *DeepSentiPeer* can extract useful features in close correspondence to human judgments. From Figure 5.3 and Table 5.4, we see that the sentiment activations are strongly correlated (negatively) with the actual and predicted recommendation scores. Therefore, we hypothesize that our model draws considerable strength if the review text has proper sentiment embedded in it. To further investigate this, we sample the papers/reviews from the ICLR 2017 test set. We consider actual review text and the sentiment embedded therein to examine the performance of the system (See Table 5.5). We truncate the lengthy review texts and provide the OpenReview links for reference. Figure 5.5 shows the heatmaps of Vader sentiment scores generated for individual sentences corresponding to each paper review in Table 5.5. We hereby acknowledge that since the scholarly review texts are mostly objective and not straightforward, the score for neutral polarity is strong as opposed to positive, and negative. But still, we can see visible polarities for review sentences which are positive or negative in sentiment. For instance, the second last sentence(s9): “*The paper is not well written either*” from R1 has visible negative weight in the heatmap (Figure 5.5). Same can be observed for the other review sentences as well.

---

<sup>4</sup>We call them as Sentiment Activations

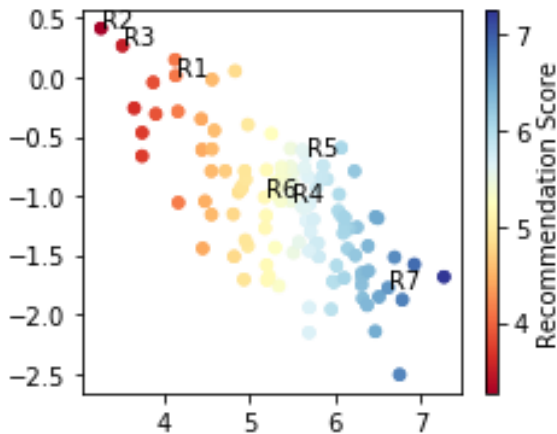


Figure 5.3: Projections of the output activations of the final layer of *MLP\_Senti*. Points are annotated for Reviews from Table 5.5. X: Predicted Recommendation Scores, Y: Sentiment Activations

Table 5.4: Pearson Correlation (PC) Coefficient between the *Recommendation Scores* and *Sentiment Activations*. This is to account for the fact that sentiment is actually correlated with the prediction signifying the strength of the model.

Scores	PC
Actual vs Prediction	0.97
Prediction vs Sentiment Activations	-0.93
Actual vs Sentiment Activations	-0.91

Table 5.5: A qualitative study of the effect of sentiment in the overall recommendation score prediction. Prediction  $\rightarrow$  is the overall recommendation score predicted by our system, Actual  $\rightarrow$  is the recommendation score given by reviewers. **Senti\_Act** are the output activations from the final layer of *MLP\_Senti* which are augmented to the decision layer for final recommendation score prediction. The correspondence between the sentiment embedded within the review texts and Sentiment Activations are fairly visible in Figure 5.3. Kindly refer to Figure 5.5 for polarity strengths in individual review sentences. The OpenReview links in the table above give the full review texts.

#	Paper Title	Review Text	Prediction	Actual	Senti_Act
---	-------------	-------------	------------	--------	-----------

## 5.2 Predicting Peer Review Outcome

R1	Multi-label learning with the RNNs for Fashion Search	—The technical contribution of this paper is not clear. Most of the approaches used are standard state-of-art methods and there are not much novelties. For a multi-label recognition task, there are other available methods, e.g. using binary models, changing cross-entropy loss function, etc. There is not any comparison between the RNN method and other simple baselines. The order of the sequential RNN prediction is not clear either. It seems that the attributes form a tree hierarchy, and that is used as the order of sequence. The paper is not well written either.— <a href="https://openreview.net/forum?id=HyWDCXjg&amp;noteId=B1Mp8grVl">https://openreview.net/forum?id=HyWDCXjg&amp;noteId=B1Mp8grVl</a>	4	3	0.01
R2	Transformation-based Models of Video Sequences	—While I agree with the authors on these points, I also find that the paper suffer from important flaws. Specifically: - the choice of not comparing with previous approaches in term of pixel prediction error seems very "convenient", to say the least. While it is clear that the evaluation metric is imperfect, it is not a reason to completely dismiss all quantitative comparisons with previous work. The frames output by the network on, e.g. the moving digits datasets (Figure 4), looks ok and can definitely be compared with other papers. Yet, the authors chose not to, which is suspicious.— <a href="https://openreview.net/forum?id=HkzAAvc&amp;noteId=SJE7-lkVx">https://openreview.net/forum?id=HkzAAvc&amp;noteId=SJE7-lkVx</a>	3	3	0.41
R3	Efficient Calculation of Polynomial Features on Sparse Matrices	—Many more relevant papers should be cited from the recent literature. The experiment part is very weak. This paper claims that the time complexity of their algorithm is $O(d^k D^k)$ , which is an improvement over standard method $O(d^k)$ by a factor $d^k$ . But in the experiments, when $d=1$ , there is still a large gap ( 14s vs. 90s) between the proposed method and the standard one. The authors explain this as "likely a language implementation", which is not convincing. To fairly compare the two methods, of course you need to implement both in the same programming language and run experiments in the same environment. For higher degree feature expansion, there is no empirical experiments to show the advantage of the proposed method.— <a href="https://openreview.net/forum?id=S1j4RqYxg&amp;noteId=B17Fn04Vg">https://openreview.net/forum?id=S1j4RqYxg&amp;noteId=B17Fn04Vg</a>	4	3	0.27

R4	Efficient Vector Representation for Documents through Corruption	—While none of the pieces of this model are particularly novel, the result is an efficient learning algorithm for document representation with good empirical performance. Joint training of word and document embeddings is not a new idea, nor is the idea of enforcing the document to be represented by the sum of its word embeddings (see, e.g. see, e.g. "The Sum of Its Parts": Joint Learning of Word and Phrase Representations with Autoencoders' by Lebet and Collobert). Furthermore, the corruption mechanism is nothing other than traditional dropout on the input layer. Coupled with the word2vec-style loss and training methods, this paper offers little on the novelty front. On the other hand, it is very efficient at generation time, requiring only an average of the word embeddings rather than a complicated inference step as in Doc2Vec. Moreover, by construction, the embedding captures salient global information about the document – it captures specifically that information that aids in local-context prediction. For such a simple model, the performance on sentiment analysis and document classification is quite encouraging. Overall, despite the lack of novelty, the simplicity, efficiency, and performance of this model make it worthy of wider readership and study, and I recommend acceptance.— <a href="https://openreview.net/forum?id=B1Igu2ogg&amp;noteId=rJBM9YbVg">https://openreview.net/forum?id=B1Igu2ogg&amp;noteId=rJBM9YbVg</a>	6	7	-1.04
R5	R5 Towards a Neural Statistician	—Hierarchical modeling is an important and high impact problem, and I think that it's under-explored in the Deep Learning literature. Pros:-The few-shot learning results look good, but I'm not an expert in this area.-The idea of using a "double" variational bound in a hierarchical generative model is well presented and seems widely applicable. Questions:-When training the statistic network, are minibatches (i.e. subsets of the examples) used?-If not, does using minibatches actually give you an unbiased estimator of the full gradient (if you had used all examples)? For example, what if the statistic network wants to pull out if *any* example from the dataset has a certain feature and treat that as the characterization. This seems to fit the graphical model on the right side of figure 1. If your statistic network is trained on minibatches, it won't be able to learn this characterization, because a given minibatch will be missing some of the examples from the dataset. Using minibatches (as opposed to using all examples in the dataset) to train the statistic network seems like it would limit the expressive power of the model— <a href="https://openreview.net/forum?id=HJDBUF5le&amp;noteId=HyWm1orEx">https://openreview.net/forum?id=HJDBUF5le&amp;noteId=HyWm1orEx</a>	6	8	-0.65



## 5.2 Predicting Peer Review Outcome

R6	A recurrent neural network without chaos	The authors of the paper set out to answer the question whether chaotic behaviour is a necessary ingredient for RNNs to perform well on some tasks. For that question's sake, they propose an architecture which is designed to not have chaos. The subsequent experiments validate the claim that chaos is not necessary. This paper is refreshing. Instead of proposing another incremental improvement, the authors start out with a clear hypothesis and test it. This might set the base for future design principles of RNNs. The only downside is that the experiments are only conducted on tasks which are known to be not that demanding from a dynamical systems perspective; it would have been nice if the authors had traversed the set of data sets more to find data where chaos is actually necessary. <a href="https://openreview.net/forum?id=S1dIzvclg&amp;noteId=H1LYxY84l">https://openreview.net/forum?id=S1dIzvclg&amp;noteId=H1LYxY84l</a>	5	8	-1.01
R7	Batch Policy Gradient Methods for Improving Neural Conversation Models	The author propose to use a off-policy actor-critic algorithm in a batch-setting to improve chat-bots. The approach is well motivated and the paper is well written, except for some intuitions for why the batch version outperforms the on-line version (see comments on "clarification regarding batch vs. on-line setting"). The artificial experiments are instructive, and the real-world experiments were performed very thoroughly although the results show only modest improvement. <a href="https://openreview.net/forum?id=rJfMusFl1&amp;noteId=H1bSmr4x">https://openreview.net/forum?id=rJfMusFl1&amp;noteId=H1bSmr4x</a>	7	7	-1.77

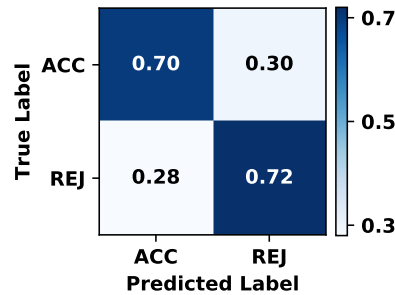


Figure 5.4: Normalized Confusion Matrix for Accept/Reject Decisions on ICLR 2017 test data with *DeepSentiPeer*(Paper+Review+Sentiment) model.

Besides the objective evaluation of the paper in the peer reviews, the reviewer's opinion in the peer review text holds strong correspondence with the overall recommendation score. We can qualitatively see that the reviews R1, R2, and R3 are polarized towards the negative

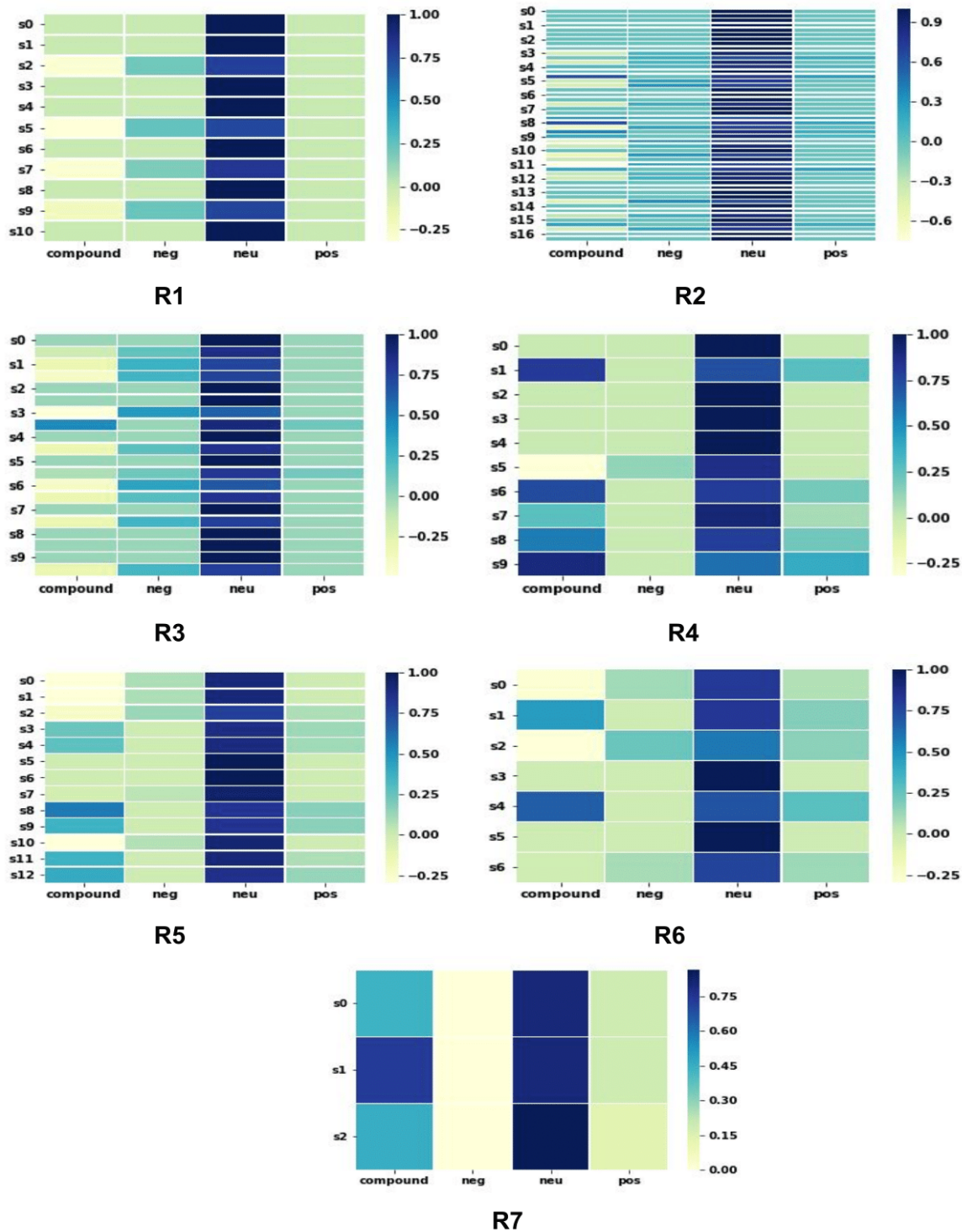


Figure 5.5: Heatmaps of the sentence-wise VADER sentiment polarity of reviews considered in Table 5.5. Reviews generally reflect the polarity of the reviewer towards the respective work.  $s_0 \dots s_n \rightarrow$  are the sentences in the peer review texts.

## 5.2 Predicting Peer Review Outcome

---

sentiment (Table 5.5). Our model can efficiently predict a reasonable recommendation score with respect to human judgment. Same we can say for R7 where the review mostly signifies a positive sentiment polarity. R6 provides an interesting observation. We see that the review R6 is not very expressive for such a high recommendation score 8. It starts with introducing the authors work and listing the strengths and limitations of the work without much (and necessary) details. Our model hence predicts 5 as the recommendation score. Whereas R4 can be seen as the case of a usual well-written review, expressing the positive and negative aspects of the paper coherently. Our model predicts 6 for an actual recommendation score of 7. These validate the role of the reviewer’s opinion and sentiment to predict the recommendation score, and our model is competent enough to take into account the overall polarity of the review-text to drive the prediction. Figure 5.4 presents the confusion matrix of our proposed model on ICLR 2017 test data for Task 2.

### 5.2.6 Conclusion

Here in this investigation, we show that the reviewer sentiment information embedded within peer review texts could be leveraged to predict the peer review outcomes. Our deep neural architecture makes use of three information channels: the paper full-text, corresponding peer review texts and the sentiment within the reviews to address the complex task of decision making in peer review. With further exploration, we aim to mould the ongoing research to an efficient AI-enabled system that would assist the journal editors or conference chairs in making informed decisions. However, considering the sensitivity of the topic, we further dive deep into exploring the subtle nuances that leads into the grading of peer review aspects. We found that review reliability prediction should prelude these tasks since not all reviews are of equal quality or are significant to the final decision making.

### 5.3 Peer Review Significance

The peer-review system is considered as the *holy grail* of scientific research validation. The recent unprecedented growth in paper submissions in major Machine Learning and AI conferences is placing a grand challenge to the community to maintain the high-quality reviewing practices while also ensuring fair evaluation of the manuscripts. Add to that, the long-standing pressing problems like arbitrariness, biases, and inconsistencies in peer reviews pose additional challenges due to the increase in reviewing load and shortage of experienced reviewers. Most of the proposed measures to counter this problem focus on mentorship and training programs for human evaluators. We argue that automatically adjudging peer-review quality can help academia penalize poor reviews, recognize and incentivize serious reviewers, and encourage fair decision-making. Here in this work, we propose the first attempt to quantify the *quality of a peer review* and develop a deep computational approach to predict the same. We hypothesize *exhaustiveness* and *strength* as two central aspects of a good quality peer review. Our deep neural network trained in a multi-task setting with scaffolds each for *exhaustiveness* and *strength* achieve encouraging performance over standard baselines.

Peer review is the foremost system of scientific research validation and, at the same time, the measure of the quality of scientific progress. However, this paper-vetting system is not without flaws. There are studies highlighting the bias [223], inconsistencies inconsistency [224, 221], arbitrariness [225], etc. as exhibited by the peer-review reports. The fate of the research work decided based on these *flawed* review reports has the potential to degrade the trustworthiness and the integrity of the system, resulting in good research getting ignored and sub-par papers getting approval. One such glaring example is the recent retractions of refereed COVID-19 research articles [226]. Area chairs/editors have long been hauled to mitigate such issues by assigning expert reviewers and evaluating reviewers' comments to make an informed decision about the research work. However, the exponential rise in paper submissions in recent years has put the mechanism of the peer-review system under serious stress. Chairs and Editors are now experiencing a dearth of experienced reviewers, sometimes delegating this critical job to inexperienced, non-expert reviewers. This stress has led the researchers argue for alternatives for review mechanism (e.g., Association for Computational Linguistics(ACL) conference proposing new guidelines<sup>5</sup>, Empirical Methods in Natural Language Processing (EMNLP) 2020 conference releasing strict rubrics for research evaluation<sup>6</sup>, AAAI 2021 having AI for conference management sub-track), while also organizing mentorship programs for reviewers [227], [228]. These efforts

---

<sup>5</sup><https://acl2020.org/reviewers/>

<sup>6</sup><https://2020.emnlp.org/blog/2020-05-17-write-good-reviews>

### 5.3 Peer Review Significance

---

are commendable to restore faith in the widely accepted method of scholarly communication. However, the peer-review system also suffers from yet another major issue, i.e., reviewers' bias to invest less time in this critical job [229]. Sometimes the reviewers themselves are not sure with their judgment on the merit of the work, which becomes evident with their reviews [230] [231]. Sometimes reviewers also evaluate the submitted research based on some poor indicators (e.g., whether the work is state-of-the-art or not [232]). Thus, the majority of reform-programs targeted towards reviewers might not be sufficient. Furthermore, given the voluntary nature of the job and stringent timelines, the quality of the reviews is affected many a time. It is relatively not uncommon that many authors are left dissatisfied after a journal or conference decision notification not only because of the negative outcome but also because the review was not detailed or not constructive enough to highlight the deficiencies of the work under scrutiny. This could be frustrating for an enthusiastic prospective author, especially in the double-blind conference model without a rebuttal period. Peer reviews are often looked upon as the community's constructive feedback, a way to identify the weakness of the research to mitigate next, and move forward [233]. Although there had been numerous community efforts towards writing good quality reviews [234], sadly, current academia is still struggling with this pressing issue.

We believe that an important direction towards re-establishing trust in the peer-review system should be an attempt to establish trust in the peer-review reports. Therefore, promoting quality and objective review reports should also carry a system to detect and penalize *flawed* review reports. Given the submission load and the urgency of this direction, an automatic system to judge reviews for quality would largely help. Recent exciting developments in Natural Language Processing can be leveraged to define and detect the surface level indicators of *flawed* review text. In this work, we propose such a system. Our system takes a review and grades it for its *quality*. There are numerous definitions what defines the *quality of a review* [235]. We assert that: *a good review should comment on the important sections (e.g., Methodology, Experiments, etc.) and address the critical aspects of the paper (e.g., Novelty, Theoretical Soundness, etc.) while clearly bringing out the reviewer's stand on the work..* We, therefore, define *quality* in two components: *Exhaustiveness* and *Strength*. Our deep neural architecture takes as input full-paper text, full review-text and is trained in a multi-task setting. We use review confidence score prediction and review recommendation score as scaffold tasks for *Exhaustiveness* and *Strength* respectively.

Our work in this direction is in line with the recent efforts to incorporate Artificial Intelligence (AI) in the peer review pipeline [48, 236, 72]. Publishers and allied stakeholders are already considering AI-assistants in the peer-review cycle [237]. An AI that could grade the reviews

based on some *quality* standards would help regulate a thrust on poor reviews. It would also help human reviewers take up their job seriously. Editors would discard trivially written reviews and ask emergency reviewers to step in. It could also be a step towards building a reviewer profile data of their review quality, thereby increasing trust in reviewers.

**Our contributions in this work are as follows:**

1. We define *quality of a review* in a systemic way in terms of two components: *Exhaustiveness* and *Strength*.
2. We propose a scoring mechanism and grade 1002 reviews from the ICLR-2018 conference for our defined scoring criteria.
3. We propose a deep neural architecture to predict the *quality of a review* using scaffolds for *Exhaustiveness* and *Strength* in a multi-task setting. We define review confidence score prediction and review recommendation score prediction as scaffold tasks for *Exhaustiveness* and *Strength* respectively.

### 5.3.1 Quality of a Review

Peer review reports present an objective evaluation of the submitted work for research validation. They discuss the merit of the work, its limitations, strengths, etc., while bringing out the reviewer’s opinion about the quality of the work. However, we can all agree with the proposition that in the present day scenario, not all peer reviews are equal [238]. Often, the reviews are written in haste or by amateur reviewers and do not contain enough logical narrative or details to consider in the final decision. Sometimes the reviewers themselves are not sure with their judgment on the merit of the work, and that becomes evident with their reviews [230, 231]. To formulate this challenging and subjective problem of defining significant peer reviews, we propose that a significant review may consist of two crucial components, *Exhaustiveness* and implicit *Strength*. We view *Exhaustiveness* as to how detailed the review is and *Strength* as how opinionated it is in reflecting the reviewer’s perspective. As per the rubrics defined in [235], we expect a detailed review to be demonstrating an understanding of the work with a concise summary, evaluating the writing quality, evaluating the novelty, should comment on the experimental section, results, and finally should critique on major components of the paper. We envisage review as a kind of knowledge derived from the paper (combined with reviewer pragmatics and domain-expertise), and as such, there should be visible overlap between the two. The review text should also manifest the reviewer’s opinion as the review is supposed to be a critique of the paper. A review may give a strong opinion about a paper without giving much explanation. A good review should be detailed and also high on the opinion. We want to distinguish reviews

### 5.3 Peer Review Significance

written by experienced reviewers vs. reviews written by novice, non-experts, or slack reviewers. And we hypothesize a good review should be comprehensive enough to discuss the merit of the work based on aspects like Impact, Empirical Soundness, etc., and opinionated enough to bring out the reviewer opinion about them. However, we also maintain that the descriptiveness and opinionatedness of the review can reflect in different amounts across different reviews (See Table 5.6). For example, when the work is very excellent, we expect to see not many comments about the work in general, but the opinion will be strongly positive. Thus, we define three different rubrics for review quality estimation: 1) *Review Exhaustiveness* 2) *Reviewer Aspect* 3) *Review Intensity*.

Table 5.6: Example reviews illustrating *Exhaustiveness* and the *Intensity* components of our definition of the *Quality* of an academic peer-review. Note that these two components manifest differently in different reviews.

Review Text	Comment
<b>R1:</b> The paper is relatively clear to follow, and implement. The main concern is that this looks like a class project rather than a scientific paper. For a class project, this could get an A in a ML class! In particular, the authors take an already existing dataset, design a trivial convolutional neural network, and report results on it. There is absolutely nothing of interest to ICLR except for the fact that now we know that a trivial network is capable of obtaining 90% accuracy on this dataset. <a href="https://openreview.net/forum?id=SyhcXjy0Z&amp;noteId=Hk2HjIfxG">https://openreview.net/forum?id=SyhcXjy0Z&amp;noteId=Hk2HjIfxG</a> [Recommendation Score: 1]	Though the review says strongly about the quality of the work but doesn't go into the specific details where the paper lacks. <b>Only Strength</b>
<b>R2:</b> This paper introduces an architecture for training a MT model without any parallel material, and tests it on benchmark datasets (WMT and captions) for two language pairs. Although the resulting performance is only about half that of a more traditional model, the fact that this is possible at all is remarkable. The method relies on fairly standard components which will be familiar to most readers: a denoising auto-encoder and an adversarial discriminator. Not much detail is given on the actual models used, for which the authors mainly refer to prior work. This is disappointing: the article would be more self-contained by providing even a high-level description of the models, such as provided (much too late) for the discriminator architecture. Misc comments: "domain" seems to be used interchangeably with "language." This is unfortunate as "domain" has another, specific meaning in NLP in general and SMT in particular. Is this intentional (if so what is the intention?) or is this just a carry-over from other work in cross-domain learning? Section 2.3: How do you sample permutations for the noise model, with the constraint on reordering range, in the general case of sentences of arbitrary lengths? Section 2.5: "the previously introduced loss [...] mitigates this concern" – How? Is there a reference backing this? Figure 3: In the caption, what is meant by "(t) = 1"? Are these epochs only for the first iteration (from M(1) to M(2))? Section 4.1: Care is taken to avoid sampling corresponding src and tgt sentences. However, was the parallel corpus checked for duplicates or near duplicates? If not, "aligned" segments may still be present. (Although it is clear that this information is not used in the algorithm) This yields a natural question: Although the two monolingual sets extracted from the parallel data are not aligned, they are still very close. It would be interesting to check how the method behaves on really comparable corpora where its advantage would be much clearer. Section 4.2 and Table 1: Is the supervised learning approach trained on the full parallel corpus? On a parallel corpus of similar size? Section 4.3: What are the quoted accuracies (84.48% and 77.29%) measured on? Section 4.5: Experimental results show a regular improvement from iteration 1 to 2, and 2 to 3. Why not keep improving performance? Is the issue training time? References: (He, 2016a/b) are duplicates. <a href="https://openreview.net/forum?id=rkYTTf-AZ&amp;noteId=r1uaaZRxf">https://openreview.net/forum?id=rkYTTf-AZ&amp;noteId=r1uaaZRxf</a> [Recommendation Score: 7]	The review starts with summarizing the paper, talks about technical aspects in much detail, and also holds opinion of the reviewer about several paper aspects. Overlap with the paper is higher than R1. <b>Exhaustiveness + Strength</b> manifested.

Though many definitions can be proposed for a subjective measure like *quality*, our definition conforms to the objective of detecting non-informative reviews, which, unfortunately, are very

common in academia. To this end, our *Review Exhaustiveness* component serves as a simple heuristic for whether the reviewer has read/understood the paper. If the reviewer has read/understood the paper, we assert that the review would be more detailed with comments on several parts of the paper, be it the problem statement, the proposed methodology, or the presented results, etc. This measure works against the kind of reviews as one illustrated in R1 and R2 below.

**R1**

This paper analyzes the expressiveness and loss surface of deep CNN. I think the paper is clearly written, and has some interesting insights.

*ICLR 2018, id=id=BJjqybcW*

**R2**

I find this paper not suitable for ICLR. All the results are more or less direct applications of existing optimization techniques, and not provide fundamental new understandings of the learning REPRESENTATION.

*ICLR 2018, id=BJDEbngCZ*

Another important role of the review is to assess the impact of the work in broader academic field. Reviewers assess proposed works for its novelty, soundness of the research methodology, etc. *Reviewer aspects* score puts thrust on this important role of an academic review. A review should have information which not only discuss the research but also validate it. Note that we split the descriptiveness of the review in two components: *Review Exhaustiveness* and *Reviewer aspect* scores. The third component to the definition Review Strength is a basic measure of review opinionatedness, and as such quantify the intensity of the review. With intensity, we do not mean the positive or negative sentiment of the review, but how strong a particular sentiment is. We now describe each of the components of our scoring mechanism:

## Review Exhaustiveness

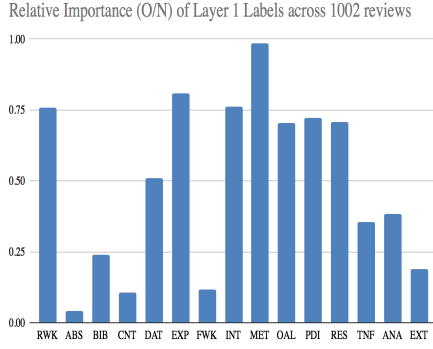
We define  $E = \{Abstract(ABS), Introduction(INT), Related\ Works(RWK), Problem\ Definition/Idea(PDI), Data/Datasets(DAT), Methodology(MET), Experiments(EXP), Results(RES), Tables\ and\ Figures(TNF), Analysis(ANA), Future\ Work(FWK), Overall(OAL), Bibliography(BIB), External(EXT)\}$ . Considering  $E$ , we score review exhaustiveness for review  $r$  as follows:

$$S_e = \frac{\left(\sum_{l_i \in E} w_{l_i} f_{l_i}\right) c n_s}{\sum_{l_i \in E} f_{l_i}} \quad (5.12)$$

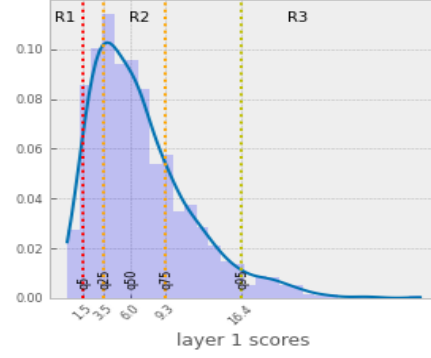
where,  $w_{l_i}$  and  $f_{l_i}$  denote the relative importance and the frequency of the label  $l_i \in E$  in  $r$ .  $n_s$  equals total number of sentences in the review, and  $c$  denotes the coverage of the review defined



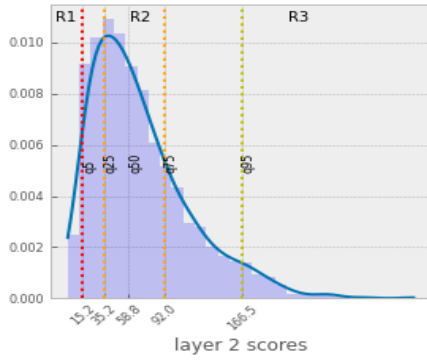
### 5.3 Peer Review Significance



(a) Relative presence of section labels in our dataset views following a ND curve



(b) Exhaustiveness scores across the annotated re-



(c) Reviewer aspect scores (reviewer subjectivity) across the annotated reviews following a ND curve

Figure 5.6: a) Relative importance of labels for Review Exhaustiveness (Layer 1). b) Distribution of the *Exhaustiveness* scores for the annotated reviews. c) Distribution of the *Reviewer Aspect* scores for the annotated reviews (Layer 2)

as

$$c = \frac{|\{l \in E, f_l \neq 0\}|}{|E|} \quad (5.13)$$

We consider relative importance in our scoring function to give more importance to certain sections, like Methodology, Experiments, etc. which are discussed in almost all the reviews (See Fig. 5.6a). This is intuitive as ICLR being an empirical venue, it's imperative for the reviewers to talk about methodology, experiments, related work for comparison, etc. And as such we agree that the scoring mechanism depends on the venue (whether it's a theoretical venue or empirical venue). The *Exhaustiveness* scores follow a normal distribution as shown in Fig 5.6b.

#### Reviewer Subjectivity

Review texts also evaluate the submitted manuscript base on other broader aspects like impact, novelty, appropriateness, etc. We hereby define  $S = \{Appropriateness (APR), Originality/Nov-$

elty (NOV), Significance/Impact (IMP), Meaningful Comparison (CMP), Presentation and Formatting (PNF), Recommendation (REC), Empirical/Theoretical Soundness (EMP), Substance (SUB), Clarity(CLA)} as the set containing this reviewer subjectivity aspects. We again score a review  $r$  for reviewer subjectivity as:

$$S_s = \frac{\left(\sum_{l_i \in S} w_{l_i} f_{l_i}\right) cn_s}{\sum_{l_i \in E} f_{l_i}} \quad (5.14)$$

Symbols carry the same meaning as before. These scores also follow a normal distribution (see Fig. 5.6c).

## Review Strength

We define review strength as the intensity of opinion (Positive, Negative, Neutral) as expressed in the review. We use Sentiment Intensity Pipeline from Huggingface [239] to get the sentiment intensity of each of the sentences in the review, and then we average them to define the intensity of the review. The intensity values lie in the range  $[-1, 1]$ .

### 5.3.2 Dataset Description

To proceed with our investigation, we require the papers, reviews with their confidence score and recommendation score. To eliminate the bias, we also need rejected papers and their reviews. However, peer review texts are sensitive and not straightforward to obtain, especially the rejected papers' data. We collect papers, reviews, recommendation scores, and confidence scores corresponding to the three editions of the International Conference on Learning Representations (ICLR) from OpenReview<sup>7</sup>. We take only the official ICLR appointed reviewer comments<sup>8</sup> for our experiments. We also study a subset of these reviews and score these reviews based on our scoring mechanism. The dataset details are in Table 5.7.

Table 5.7: Dataset Statistics

Venues	#Papers	#Reviews	Acc/Rej
ICLR 2017	427	1281	172/255
ICLR 2018	909	2741	336/573
ICLR 2019	1418	4254	500/918
Total	2754	8276	1008/1746

<sup>7</sup><https://openreview.net>

<sup>8</sup>three reviews per paper

### 5.3.3 Methodology

Our initial analysis shows an interplay between the recommendation score, confidence score, review sentiment, and type of review. Some key observations are:

- When the recommendation score, confidence score and the review sentiment (positive) are high, the number of annotations  $\sum_{l_i}^{|E|} f_{l_i}$  from Eq. 5.14 is low (sections/aspects). This seems intuitive, as when the paper is extremely good, there really isn't much to talk about.
- When the recommendation score, confidence score is high, and the review sentiment (negative) is high, the number of annotations (sections/aspects) is again low. This also makes sense, as when the paper is extremely bad, reviewers might struggle where to even begin with.
- One interesting trend is when the recommendation score is high, and the sentiment is positive, but the confidence score is relatively low (2-3). Even in this case, the total number of annotations is less (i.e., review text corresponding to limited sections or aspects). One possible explanation of this could be that non-expert reviewers write reviews without understanding the work in depth.

Given these kind of trends, it is clear that there is some interplay between recommendation score, confidence score and the quality scores. Given less training data for the task, it seems promising to exploit these relationships for the prediction. Sadly, these scores are not without flaws and can hint non-desirable relationships, as evident from the third observation above. However, we can hope that such noisy trends will nullify if we average over the larger dataset. Thus, we formulate recommendation score and confidence score also as a learning problem to get away with the flaws in the original scores. We hypothesize that a machine learning model trained to quantify the subjectivity of the reviews would give fair and objective scores to the reviews. And then we exploit the interplay between the recommendation score, confidence score and the quality scores in a multi-task framework. Our model takes a paper and a review as input and train it by predicting the confidence, recommendation, and the quality scores. Our *main task* is to predict the quality scores, and our *scaffold tasks* are recommendation score prediction and confidence score prediction. We describe the main components of our model in the following sections:

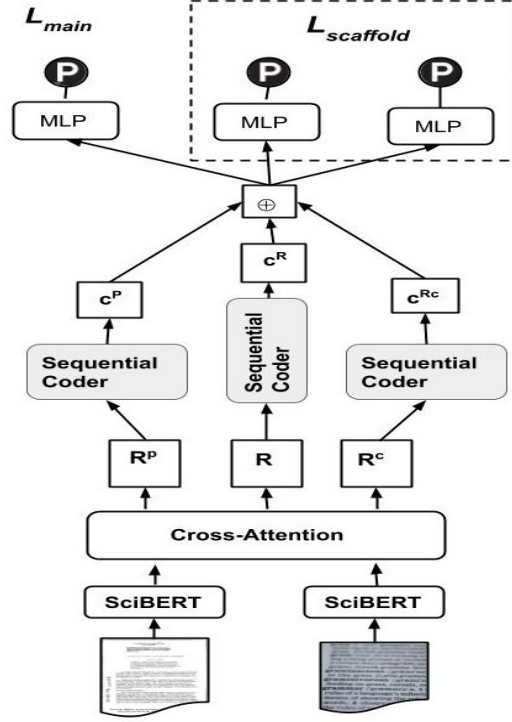


Figure 5.7: Our architecture for predicting the significance of peer reviews (main task). The model is trained on the peer review decisions (secondary task)

## Pre-processing

For information extraction from the papers, we first convert the PDF's into .json using the Science Parse<sup>9</sup> library. We do not consider the figures, tables in the processing as these entities are not correctly parsed. We consider only paper full-text sentences and strip of the headings. Processing the reviews were pretty straightforward, and we strip all the additional information and consider only the review texts in further processing.

## Encoder

The inputs to our model are full paper and review text. We denote the paper

$$P = (s_1^p, s_2^p, \dots, s_{n_p}^p) \quad (5.15)$$

and review  $R = (s_1^r, s_2^r, \dots, s_{n_r}^r)$  as the collection of their respective sentences. For a sentence  $s_i^p$  we get a  $d$  dimensional embedding vector  $\mathbf{s}_i^p \in \mathbb{R}^d$  using SciBERT encoder [240, 241]. We use the pretrained SciBERT model to incorporate rich representation power, as this language model is trained on large scientific corpora. We get the paper representation  $\mathbf{P} \in \mathbb{R}^{n_p \times d}$  by concatenating

<sup>9</sup><https://github.com/allenai/science-parse>

### 5.3 Peer Review Significance

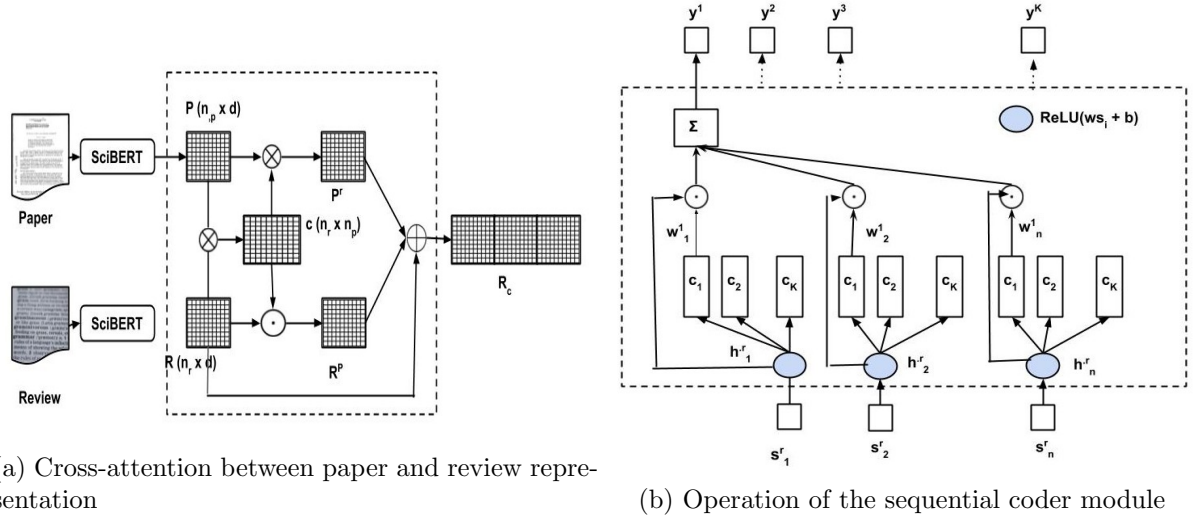


Figure 5.8: The cross-attention and sequential coder module from Figure 5.7

these vectors.

$$\mathbf{P} = \mathbf{s}_1^p \oplus \mathbf{s}_2^p \oplus \dots \oplus \mathbf{s}_{n_p}^p, \mathbf{P} \in \mathbb{R}^{n_p \times d} \quad (5.16)$$

Similarly, we get the review representation  $\mathbf{R} \in \mathbb{R}^{n_r \times d}$  as

$$\mathbf{R} = \mathbf{s}_1^r \oplus \mathbf{s}_2^r \oplus \dots \oplus \mathbf{s}_{n_r}^r, \mathbf{R} \in \mathbb{R}^{n_r \times d} \quad (5.17)$$

#### Context Modeling

We employ the basic co-attention module [242] to extract the relative representation of paper and review with respect to each other (See Figure 5.8a).

We get the affinity matrix  $\mathbf{E} \in \mathbb{R}^{n_r \times n_p}$  as follows:

$$\mathbf{E}_{ij} = \frac{1}{\sqrt{d}} \mathcal{F}\{(\mathbf{s}_i^r)^T\} \mathcal{F}\{\mathbf{s}_j^p\}, \mathbf{E}_{ij} \in \mathbb{R} \quad (5.18)$$

Here  $\mathcal{F}$  is a linear layer ( $\mathbf{w}^T \mathbf{x} + \mathbf{b}$ ). Thus,  $\mathbf{E}_{ij}$  gives the measure of similarity between the sentence  $s_i^r$  and  $s_j^p$ , which we convert to attention weights using the following normalizations:

$$c_{ij} = \frac{\exp \mathbf{E}_{ij}}{\sum_{k=1}^{n_p} \exp \mathbf{E}_{ik}}, c_{ij} \in \mathbb{R} \quad (5.19)$$

$$\hat{c}_{ij} = \frac{\exp \mathbf{E}_{ij}}{\sum_{k=1}^{n_r} \exp \mathbf{E}_{kj}}, \hat{c}_{ij} \in \mathbb{R} \quad (5.20)$$

We get the relative representation of review sentence  $\mathbf{s}_i^r$  with respect to paper  $P$  by

$$\mathbf{r}_i^p = \sum_{j=1}^{n_p} c_{ij} \mathbf{s}_i^r, \mathbf{r}_i^p \in \mathbb{R}^d \quad (5.21)$$

We further get the review representation in light of the paper sentence  $\mathbf{s}_i^p$  as follows:

$$\hat{\mathbf{r}}_i^p = \sum_{j=1}^{n_r} \hat{c}_{ij} \mathbf{s}_j^r, \hat{\mathbf{r}}_i^p \in \mathbb{R}^d \quad (5.22)$$

Similarly, we get the paper representation in view of the review sentence  $\mathbf{s}_i^r$  by:

$$\mathbf{p}_i^r = \sum_{j=1}^{n_p} c_{ij} \mathbf{s}_j^p, \mathbf{p}_i^r \in \mathbb{R}^d \quad (5.23)$$

We concatenate all the corresponding vectors in  $\mathbf{r}_i^p, \hat{\mathbf{r}}_i^p, \mathbf{p}_i^r$  to get  $\mathbf{R}^p \in \mathbb{R}^{n_r \times d}, \hat{\mathbf{R}}^p \in \mathbb{R}^{n_p \times d}, \mathbf{P}^r \in \mathbb{R}^{n_r \times d}$ .

### Sequential Feature Extractor

Now we do the individual processing of the  $\mathbf{R}^p, \mathbf{R}$ , and  $\mathbf{R}^c = \mathbf{R}^p \oplus \mathbf{P}^r \oplus \mathbf{R}$  in our *coder* module (Figure 5.8b). In simple terms, a *coder* module is a collection of  $K$  attention modules. To encode extract useful information from the review  $\mathbf{R}$ , a coder module can be described in terms of individual sentences of the reviews  $\mathbf{s}_1^r, \mathbf{s}_2^r, \dots, \mathbf{s}_{n_r}^r$ . The  $k^{th}$  code extraction can be done as

$$\mathbf{h}_i^r = \text{ReLU}(\mathbf{w}^T \mathbf{s}_i^r + b) \quad (5.24)$$

$$w_i^k = \frac{\exp(\mathbf{h}_i^r \cdot \mathbf{c}_k)}{\sum_{j=1}^{n_r} \exp(\mathbf{h}_j^r \cdot \mathbf{c}_k)} \quad (5.25)$$

$$y^k = \sum_{j=1}^{n_r} w_j^k \mathbf{h}_j^r \quad (5.26)$$

Thus, we get  $y^1, y^2, \dots, y^K$  features representations of a review  $R$  each encoded using a special trainable coder  $\mathbf{c}_k$  which is randomly initialized at first. This module can be seen as a shallow approximation to the popular self-attention operation in NLP. We perform this operation sequentially for  $m$  times. We get outputs  $\mathbf{c}^p, \mathbf{c}^R$  and  $\mathbf{c}^{\mathbf{R}^c}$  for each of  $\mathbf{R}^p, \mathbf{R}, \mathbf{R}^c$  respectively. Note that we are working with full paper text and full review text; thus, a shallow approximation to the self-attention technique is used. However, any feasible contextual feature representation method can be used if computation and resources are not constraints.

#### Feedforward Prediction Layers

We concatenate the outputs from the  $m^{th}$  coder layer i.e.  $\mathbf{c}^p, \mathbf{c}^R$  and  $\mathbf{c}^{R_c}$  together in one flattened vector  $\mathbf{f}_r$ . We pass  $\mathbf{f}_r$  outputs to the prediction layers for the scaffold tasks. Each task has its own prediction layers, which is a multi-layered Perceptron (MLP) layer. Thus, the last layers of the model have task-specific trainable parameters, while the parameters in the lower layers are shared.

#### Training and Experimental Setup

We describe our training routine in Algorithm 3 below. To account for the different numerical ranges of the loss values across multiple scaffold tasks, we combine recommendation task loss  $\mathcal{L}_{rec.}$  and confidence task loss  $\mathcal{L}_{conf.}$  using a method proposed in [243] to get  $\mathcal{L}_{scaffold}$  as

$$\mathcal{L}_{scaffold} = \frac{1}{2\sigma_{rec.}^2} \mathcal{L}_{rec.} + \frac{1}{2\sigma_{conf.}^2} \mathcal{L}_{conf.} + \log \sigma_{rec.} \sigma_{conf.} \quad (5.27)$$

Similarly, for  $\mathcal{L}_{main}$ ,

$$\mathcal{L}_{main} = \frac{1}{2\sigma_{exh.}^2} \mathcal{L}_{exh.} + \frac{1}{2\sigma_{subj.}^2} \mathcal{L}_{subj.} + \frac{1}{2\sigma_{int.}^2} \mathcal{L}_{int.} + \log \sigma_{exh.} \sigma_{subj.} \sigma_{int.} \quad (5.28)$$

where  $\mathcal{L}_{exh.}, \mathcal{L}_{subj.}, \mathcal{L}_{int.}$  are corresponding loss for exhaustiveness, subjectivity (aspect), and intensity respectively and  $\sigma$  is the model's observation noise parameter – capturing how much noise we have in the outputs. We also normalize all the scores so that they lie in range  $[1, 10]$ . This normalization of the *intensity* scores means that larger value indicate strong positive sentiment and lesser value indicate strong negative sentiment.

---

#### Algorithm 3 Training routine of the multi-task model

---

**Given** Main Task Dataset  $\mathcal{D}$ , Scaffold Task Dataset,  $\mathcal{D}'$ , Model  $M_\theta$ ;

**for** iter, batch ( $B_{\mathcal{D}'}$ ) in num\_iters **do**  $P, R = B_{\mathcal{D}'}$ ;  $\mathcal{L}_{scaffold} = M_\theta$ ; update  $\theta$ ;

**for** batch ( $B_{\mathcal{D}}$ ) in  $\mathcal{D}$  **do**  $P, R = B_{\mathcal{D}}$ ;  $\mathcal{L}_{main} = M_\theta$ ; update  $\theta$ ;

---

#### 5.3.4 Evaluation

As mentioned earlier, our main task is to predict the *Quality* scores for each review. This task is a regression task, so are our scaffolds. We evaluate the model on the held-out test set. Since there are no gold-standard data for evaluating the main task, we use simple baselines as comparing systems. Also, since we have very little data with gold annotated scores, we cannot compare with other sophisticated deep-learning techniques as they are data-intensive. Our major contribution

in the prediction task is using multi-task learning to facilitate the predictions even with using fewer data. We employ the following simple baselines:

1. **Mean Baseline:** As most deep neural networks models for regression tasks are susceptible to predicting the mean of the prediction target, we use this baseline as a comparing system to see if our model is not suffering from the same.
2. **Ridge Regression:** We perform simple Ridge Regression with average sentence embeddings as the input to the model. The model for Ridge Regression is a two-layer feedforward network with an L2 regularizer. We keep the model shallow and carefully tune the L2 regularizer weight to prevent overfitting (chances of which are high due to less training data).
3. **Ridge Regression with Score Fusion:** In this baseline, we additionally give recommendation score and confidence score as the input to the model, along with the average sentence embeddings as described in the previous baseline.
4. **CNN as a Sequential Coder:** We also experiment with replacing the Sequential Feature Extractor with a Convolutional Neural Network with max-pooling in the multi-task framework.

### 5.3.5 Results

Table 5.8 shows the results for all the different methods. Given the distribution of the scores (refer Fig. 5.6b and Fig. 5.6c), it's seems reasonable that the mean baseline performs better. Due to the less number of reviews with *quality* scores, ridge regression performs approximately the same as the Mean Baseline. This also means that the Ridge Regression method is under parameterized to the complexity of the task. Fusing the recommendation and confidence scores into the Ridge Regression leads to improvement, further hinting at our multi-task learning motivation. Each of the scaffolds individually also leads to good improvements to the mean baseline. Incorporating both the tasks together in a single model gives substantial improvements over the other comparing systems. Interestingly, CNN, as a sequential coder, also performs better. Furthermore, we observe that the recommendation sub-task results in good improvements in Layer 3 (*Intensity*) scores, whereas the Confidence sub-task has improved performance for Layer 1 (*Exhaustiveness*). It can be seen as an evidence that *Recommendation* scores are a proxy for *Review Intensity*, while the proxy for *Exhaustiveness* is Confidence scores. This also seems intuitive (except for reviewer's inconsistencies and implicit subjectivity in translating their views of the research work to numerical scores) as a high recommendation score means that the review text has a positive polarity (negative for otherwise). Similarly, when the reviewer has



### 5.3 Peer Review Significance

thoroughly understood/read the paper and has written the review in detail, the confidence score is expected to be high. Thus, Recommendation Score and Confidence Score prediction tasks can also be seen as scaffolds for *Exhaustiveness* and *Intensity* of the review, respectively. We hypothesize that they can take care of the inconsistency in one them by drawing information from the other task. *This is somewhat similar to area chairs/editors giving lesser importance to the reviews with high recommendation scores but lower confidence scores.*

Table 5.8: Performance of Score Prediction(Regression Task) across all the models. Layer1, Layer2, and Layer3 denotes *Exhaustive*, *Reviewer aspect*, *Intensity* tasks respectively. Training is done on ICLR 2017, 2018 and 2019 data, and testing is done on held out test data. RMSE  $\rightarrow$  Root Mean Square Error.

Model	RMSE			Cosine Similarity		
	Layer 1	Layer 2	Layer 3	Layer 1	Layer 2	Layer 3
Mean Baseline	1.551	2.110	1.675	0.893	0.870	0.908
Ridge Regression	1.556	2.091	1.649	0.898	0.881	0.910
Ridge Regression (Score Fusion)	1.553	2.172	1.604	0.901	0.888	0.915
Confidence Scaffold	1.066	1.250	1.550	0.910	0.922	0.924
Recommendation Scaffold	1.193	1.124	1.482	0.909	0.927	0.931
Both Scaffolds	0.50	0.80	1.457	0.989	0.982	0.939
CNN Sequential Coder	0.73	0.95	1.459	0.979	0.975	0.926

#### 5.3.6 Qualitative Analysis

We list some of the representative reviews to illustrate the common errors in our approach in Table 5.9. One aspect our approach lacks is getting the gold scores for *intensity*. In R1, we see that the actual *intensity* score is 8.58. However, the review text does not reflect this high positive polarity. We use a pre-trained model from Huggingface Transformer library [239] to generate the actual intensity values. Interestingly enough, our model predicts 5.55 as the score, which appears better indicator of the sentiment present in the review (if we judge qualitatively). For R2, the prediction for *reviewer aspect* score is relatively higher than the actual score. This could be due to references in the review-text, which could hint towards the Related Work (RWK) section from our layer 2 (Aspect) scoring criteria. We have relatively higher prediction for *Exhaustiveness* and *Intensity* scores. A reason for the former could be the presence of a concise summary of the paper in the beginning, and the latter can be explained by the presence of positive-sentiment attributes in the text. R4 is a detailed and descriptive review; however, it refers lesser to the paper material than its objective evaluation in a wider context. As such, the model predicts a higher *Aspect* score than the actual score (which we derive through our scoring mechanism from raw annotations).

Table 5.9: Review Significance Score Predictions

#	Review Text	Predicted Scores	Actual Score
R1	<p>This paper investigates the complexity of neural networks with piecewise linear activations by studying the number of linear regions of the representable functions. It builds on previous works of Montufar et al. (2014) and Raghu et al. (2017) and presents improved bounds on the maximum number of linear regions. It also evaluates the number of regions of small networks during training. The improved upper bound given in Theorem 1 appeared in SampTA 2017 - Mathematics of deep learning "Notes on the number of linear regions of deep neural networks by Montufar. The improved lower bound given in Theorem 6 is very modest but neat. Theorem 5 follows easily from this. The improved upper bound for maxout networks follows a similar intuition but appears to be novel. The paper also discusses the exact computation of the number of linear regions in small trained networks. It presents experiments during training and with varying network sizes. These give an interesting picture, consistent with the theoretical bounds, and showing the behaviour during training. Here it would be interesting to run more experiments to see how the number of regions might relate to the quality of the trained hypotheses. <b>Recommendation 6, Confidence 5</b></p>	<p>layer1: 2.16 layer2: 3.05 layer3: 5.55</p>	<p>layer1: 2.07 layer2: 2.78 layer3: 8.58</p>
R2	<p>The paper is grounded on a solid theoretical motivation and the analysis is sound and quite interesting. There are no results on large corpora such as 1 billion tokens benchmark corpus, or at least medium level corpus with 50 million tokens. The corpora the authors choose are quite small, the variance of the estimates are high, and similar conclusions might not be valid on a large corpus. [1] provides the results of character level language models on Enwik8 dataset, which shows regularization doesn't have much effect and needs less tuning. Results on this data might be more convincing. The results of MOS is very good, but the computation complexity is much higher than other baselines. In the experiments, the embedding dimension of MOS is slightly smaller, but the number of mixture is 15. This will make it less usable, I think it's necessary to provide the training time comparison. Finally experiments on machine translation or speech recognition should be done and to see what improvements the proposed method could bring for BLEU or WER. [1] Melis, Gbor, Chris Dyer, and Phil Blunsom. "On the state of the art of evaluation in neural language models." arXiv preprint arXiv:1707.05589 (2017). [2] Joris Pelemans, Noam Shazeer, Ciprian Chelba, Sparse Non-negative Matrix Language Modeling, Transactions of the Association for Computational Linguistics, vol. 4 (2016), pp. 329-342. [3] Shazeer et al. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. ICLR 2017.</p>	<p>layer1: 2.59 layer2: 3.25 layer3: 3.90</p>	<p>layer1: 2.31 layer2: 1.83 layer3: 5.64</p>

### 5.3 Peer Review Significance

---

<b>R3</b>	<p>Pros: The paper proposes a "bi-directional block self-attention network (Bi-BloSAN)" for sequence encoding, which inherits the advantages of multi-head (Vaswani et al., 2017) and DiSAN (Shen et al., 2017) network but is claimed to be more memory-efficient. The paper is written clearly and is easy to follow. The source code is released for duplicability. The main originality is using block (or hierarchical) structures; i.e., the proposed models split the an entire sequence into blocks, apply an intra-block SAN to each block for modeling local context, and then apply an inter-block SAN to the output for all blocks to capture long-range dependency. The proposed model was tested on nine benchmarks and achieve good efficiency-memory trade-off. Cons: - Methodology of the paper is very incremental compared with previous models.- Many of the baselines listed in the paper are not competitive; e.g., for SNLI, state-of-the-art results are not included in the paper. - The paper argues advantages of the proposed models over CNN by assuming the latter only captures local dependency, which, however, is not supported by discussion on or comparison with hierarchical CNN. - The block splitting (as detailed in appendix) is rather arbitrary in terms of that it potentially divides coherent language segments apart. This is unnatural, e.g., compared with alternatives such as using linguistic segments as blocks. - The main originality of paper is the block style. However, the paper doesn't analyze how and why the block brings improvement. -If we remove intra-block self-attention (but only keep token-level self-attention), whether the performance will be significantly worse?</p>	layer1: 3.01 layer2: 3.92 layer3: 4.20	layer1: 1.99 layer2: 2.21 layer3: 2.66
-----------	---	---	---

<b>R4</b>	<p>The paper studies the theoretical properties of the two-layer neural networks. To summarize the result, let's use the <math>\theta</math> to denote the layer closer to the label, and <math>W</math> to denote the layer closer to the data. The paper shows that a) if <math>W</math> is fixed, then with respect to the randomness of the data, with prob. 1, the Jacobian matrix of the model is full rank b) suppose that we run an algorithm with fresh samples, then with respect to the randomness of the <math>k</math>-th sample, we have that with prob. 1, <math>W_k</math> is full rank, and the Jacobian of the model is full rank. It's known (essentially from the proof of Carmon and Soudry) that if the Jacobian of the model is full rank for any matrix <math>W</math> w.r.t the randomness of the data, then all stationary points are global. But the paper cannot establish such a result. The paper is not very clear, and after figuring out what it's doing, I don't feel it really provides many new things beyond C-S and Xie et al. The paper argues that it works for activation beyond relu but result a) is much much weaker than the one with for all quantifier for <math>W</math>. result b) is very sensitive to the exactness of the events (such as <math>W</math> is exactly full rank) — the events that the paper talks just naturally never happen as long as the density of the random variables doesn't degenerate. As the author admitted, the results don't provide any formal guarantees for the convergence to a global minimum. It's also a bit hard for me to find the techniques here provide new ideas that would potentially lead to resolving this question. additional review after seeing the author's response: The author's response pointed out some of the limitation of Soudry and Carmon, and Xie et al's which I agree. However, none of this limitation is addressed by this paper (or addressed in a misleading way to some extent.) The key technical limitation is the dependency of the local minima on the weight parameters. Soudry and Carmon addresses this in a partial way by using the random dropout, which is a super cool idea. Xie et al couldn't address this globally but show that the Jacobian is well conditioned for a class of weights. The paper here doesn't have either and only shows that for a single fixed weight matrix, the Jacobian is well-conditioned. I don't also see the value of extension to other activation function. To some extent this is not consistent with the empirical observation that relu is very important for deep learning. Regarding the effect of randomness, since the paper only shows the convergence to a first-order optimal solution, I don't see why randomness is necessary. Gradient descent can converge to a first order optimal solution. (Indeed I have a typo in my previous review regarding "w.r.t. <math>k</math>-th sample", which should be "w.r.t. <math>k</math>-th update". ) Moreover, to justify the effect of the randomness, the paper should have empirical experiments. I think the writing of the paper is also misleading in several places.</p>	<p>layer1: 4.14 layer2: 5.28 layer3: 3.57</p>	<p>layer1: 3.62 layer2: 3.28 layer3: 3.17</p>
-----------	---	---	---

### 5.3.7 Ethical Issues

We understand that delegating this sensitive and task of paramount importance to scientific progress to an AI which suffers from interpretability issues, bias; reliability is not without unwarranted risks and potential misuses. Our motivation with this work is not to add algorithmic issues to the already at peril peer-review system. This work is intended as a proof-of-concept to illustrate that the most recent developments in Natural Language Processing and AI in-general can be used as an advantage to eliminate some of the bias in the peer-review process. As such, we

### 5.3 Peer Review Significance

---

want to motivate research on quantifying and judging the *quality* of peer reviews. As peer-review texts are at the harness of validating the scientific research, we want to monitor what kind of indicators are used for that validation. We ascertain that *flawed* and *unreasonable* indicators should not be the driving force of scientific progress. This work is not intended towards the quality of a paper, but on the reviews the papers received. We use the publicly available dataset from openreview.net, so no violation of confidentiality with respect to data or authors is made.

#### 5.3.8 Conclusion

Here in this work, we make a first attempt of its kind to grade peer reviews based on their exhaustiveness and strength automatically. We do not claim that our hypothesis and method addresses the entire spectrum of *peer-review quality*. We encourage further investigations on this crucial problem from an NLP/ML perspective. Obviously nature of peer-reviews would vary across domains, and it would be interesting to explore measures of exhaustiveness across disciplines other than NLP/ML.

## 5.4 Finding a Research Lineage

Finding the lineage of a research topic is crucial for understanding the prior state of the art and advancing scientific displacement. The deluge of scholarly articles makes it difficult to locate the most relevant previous work and causes researchers to spend a considerable amount of time building up their literature list. Citations play a crucial role in discovering relevant literature. However, not all citations are created equal. The majority of the citations that a paper receives provide contextual and background information to the citing papers. In those cases, the cited paper is not central to the theme of citing papers. However, some papers build upon a given paper, further the research frontier. In those cases, the concerned cited paper plays a pivotal role in the citing paper, and hence the nature of citation the former receives from the latter is *significant*. In this work, we discuss our investigations towards discovering *significant citations* of a given paper. We further show how we can leverage significant citations to build a research lineage via a *significant citation graph*. We demonstrate the efficacy of our idea with two real-life case studies. Our experiments yield promising results with respect to the current state-of-the-art in classifying *significant citations* outperforming the earlier ones by a relative margin of 20 points in terms of precision. We hypothesize that such an automated system can facilitate relevant literature discovery and help identify knowledge flow for a certain category of papers. Finding the relevant prior literature or identifying the gradual evolution of a research is critical to reviewing a new submission by a peer reviewer (if not an expert, which is highly probable in this rate of paper submissions). Hence, the current work would also assist the reviewers to locate where do the current work stands with respect to the prior relevant research.

Literature searches are crucial to discover relevant publications. The knowledge discovery that ensues forms the basis of understanding a research problem, finding the previously explored frontiers, identifying research gaps, which eventually leads to the development of new ideas. However, with the exponential growth of scientific literature (including published papers and pre-prints) [68], it is almost impossible for a researcher to go through the entire body of the scholarly works even in a very narrow domain. Citations play an important role here to find the relevant articles that further topical knowledge. However, not all citations are equally [135] effective in finding relevant research. A majority of the papers cite a work contextually [244] for providing additional background context. Such background contextual citations obviously help in the broader understanding; however, they are not central to the citing paper's theme. Some papers use the ideas in a given paper, build upon those ideas, and make displacement to the body of relevant research. Such papers are expected to acknowledge the prior work (via citing them) duly. However, the nature of citation, in this case, is different than that of the

## 5.4 Finding a Research Lineage

---

contextual citations. These citations, which heavily rely on a given work or build upon that work, are *significant citations*. However, the current citation count metric put equal weights to all the citations and is therefore not adequate to identify the papers that have significantly cited a given work and may have taken the relevant research forward. Identifying such *significant citations* are hence crucial to the literature study.

It is not uncommon that authors sometimes fail to acknowledge the role of relevant papers in stemming up their ideas [245, 246]. As a result, researchers spend a lot of their time searching for the most relevant papers to their research topic, thereby locating the subsequent papers that carried forward a given scientific idea. It is usually desirable for a researcher to understand the story behind a prior work and trace the concept’s emergence and gradual evolution through publications, thereby identifying the knowledge flow. Identifying references that are significant to a given paper and then hierarchically locating meaningful prior work is what researchers ideally do to curate their literature base.

The idea of recognizing *significant citations* is also important to understand the true impact of given research or facility. To understand how pervasive a research was in the community, it is essential to understand the influence of the given research beyond the direct citations it received. To this end, tracking *transitive* influence of a research via identifying significant citations could be one possible solution.

In this work, we develop automatic approaches to trace the lineage of given research via transitively identifying the significant citations to a given article. The overall objective of our work is two-fold:

- Accelerate relevant literature discovery via establishing a research lineage
- Find the true influence of a given work and its pervasiveness in the community beyond citation counts

There are two aspects to the problem: *identifying the relevant prior work* and *identifying the follow-up works that stemmed or are influenced by the current work*. The first aspect would facilitate relevant prior literature discovery for a paper while the second aspect would facilitate discovering knowledge flow in subsequent relevant papers. Obviously, our approach would not be a *one shoe fits for all*. Still, we believe it is effective to find investigations that build upon relevant priors, facilitate relevant literature discovery, and thereby steer towards identifying the pervasiveness of a given piece of research in the community. We base our work to classify citations

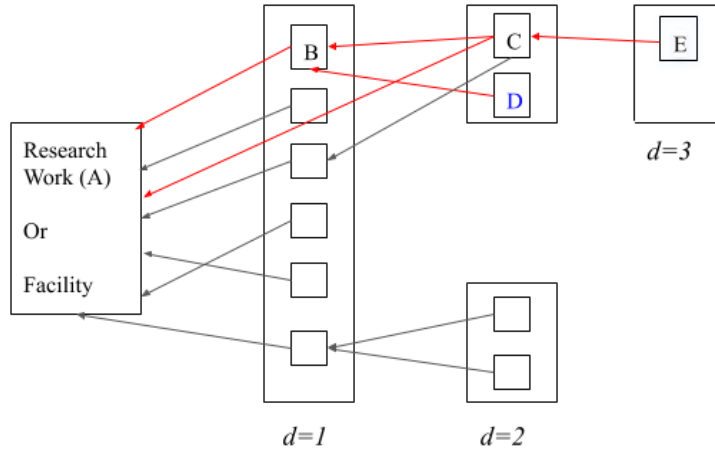


Figure 5.9: Research Lineage

as *contextual* or *significant* and trace the lineage of a research in a citation graph via identifying significant edges.

#### 5.4.1 Research Lineage

The mechanism of citations in academia is not always transparent [247, 248, 249]. Problems like coercive citations [250], anomalous citations [251], citation manipulation [252], *rich gets richer effects* [253], discriminatory citation practices [254], etc. has infested the academic community. However, in spite of all these known issues, citation counts and *h*-indices still remain the measures of research impact and tools for academic incentives, though long-debated by many [255, 256]. Usually, we measure the impact of a given paper by the direct citations it receives. However, a given research may have induced a transitive effect on other papers, which are not apparent with the current citation count measures. Figure 5.9 shows a sample citation network where A could be a paper or a research facility. We want to know how pervasive was the research or facility A in the community. At  $d=1$  are the direct citations to A. We see article B cites A significantly, or B is inspired by A. Other citations to A are background. At citation depth  $d=2$ , we see that article C and article D significantly cites B (direct citation). We see that C also cites A significantly. Finally, at citation depth  $d=3$ , E significantly cites C. We intend to understand if there is a lineage of research from A to E ( $A \rightarrow B \rightarrow C \rightarrow E$ ). Although E does not cite A directly, can we identify the influence of A on E? If E is a seminal work, receiving hundreds of citations, can we infer that A was the prior work that indirectly inspired E? We are interested in discovering such hidden inspirations to truly assess the contributions of a research article or a facility.



## 5.4 Finding a Research Lineage

---

### 5.4.2 Dataset Description

We experiment with the Valenzuela dataset [151] for our task. The dataset consists of incidental/influential human judgments on 630 citing-cited paper pairs for articles drawn from the 2013 ACL anthology, the full texts of which are publicly available. Two expert human annotators determined the judgment for each citation, and each citation was assigned a label. Using the author’s binary classification, 396 citation pairs were ranked as incidental citations, and 69 (14.3%) were ranked as influential (important) citations. For demonstrating our research lineage idea, we explore knowledge flow on our papers of *Document-Level Novelty Detection* [8] and the High Performance Computing (HPC) algorithm *MENNDL* [27]. Original authors of MENNDL helped us with manual annotation of their paper’s lineage.

### 5.4.3 Methodology

To identify significant citations, we pursue a feature-engineering approach where we curate several features from cited-citing paper pairs. The objective is to classify the citations received by a given paper into SIGNIFICANT and CONTEXTUAL. The original cited, citing papers in the Valenzuela dataset are in PDF. We convert the PDFs to corresponding XMLs using GROBID [257].

1. **Citation frequency inside the body of citing paper (F1):** We measure the number of times the cited paper is referenced from within the citing paper’s body. The intuition is that if a paper is cited multiple times, the cited paper may be significant to the citing paper.
2. **Are the authors of citing & cited paper the same? (Boolean) (F2):** We check if the authors of citing and cited paper are the same. This might be the case of self-citation or can also signal the extension of the work.
3. **Author overlap ratio (F3):** This measures number of common authors in citing and cited paper normalized to the total number of authors in citing paper. Intuition is similar to F2.
4. **Is the citation occurring in a table or figure captions? (Boolean) (F4):** The intuition is that most of the citations in tables & figures appear for comparison/significantly referencing existing work. Hence, the citing paper might be an extension of the cited article or may have compared it with earlier significant work.
5. **Is the citation occurring in groups? (Boolean) (F5):** We check if the citation is

occurring along with other citations in a group. Intuition is that such citations generally appear in related works to highlight a background detail; hence it might not be a significant citation.

6. **Number of citations to the cited paper normalized by the total number of citations made by the citing paper (F6):** This measures number of citations to the cited paper by the citing paper normalized by the total number of citation instances in the citing paper. This is to measure how frequently is the cited paper mentioned compared to other cited papers in the citing paper.
7. **Number of citations to the cited paper normalized by the total number of bibliography items in the citing paper (F7):** This measures number of citations to the cited paper normalized to the total number of bibliography items in the citing paper. Intuition is similar to F6.
8. ***tf-idf* similarity between abstracts of the cited and citing paper (F8):** We take cosine similarity between the *tf-idf* representations of the abstracts of cited and citing papers. Intuition is that if the similarity is higher, the citing paper may be inspired/extended from the cited paper.
9. ***tf-idf* similarity between titles of the cited and citing paper (F9):** We take cosine similarity between the *tf-idf* representations of the titles of cited and citing papers.
10. **Average *tf-idf* similarity between citance and abstract of the cited paper (F10):** We calculate the similarity of each citance with the abstract of the cited article and take the average of it. Citances are sentences containing the citations in the citing paper. Citances reveal the purpose of the cited paper in the citing paper. Abstracts contain the contribution/purpose statements of a given paper. Hence similarity with citances may construe that the cited paper may have been used significantly in the current paper.
11. **Maximum *tf-idf* similarity between citance and abstract of the cited paper (F11):** We take the maximum of similarity of the citances (there could be multiple citation instances of the same paper in a given paper) with the abstract of the cited paper.
12. **Average *tf-idf* similarity between citance and title of the cited paper (F12):** We calculate the similarity of each citance with the title of the cited paper and take an average of it.
13. **Maximum *tf-idf* similarity between citance and title of the cited paper (F13):** We take the maximum of similarity of the citances with the title of the cited paper.

## 5.4 Finding a Research Lineage

---

14. **Average Length of the Citance (F14):** Average length of the citances (in words) for multiple citances. Intuition is that if the citing paper has spent many words on the cited article, it may have significantly cited the corresponding article.
15. **Maximum Length of the Citance (F15):** Maximum length of the citances (in words).
16. **No. of words between citances (F16):** We take the average of the number of words between each pair of consecutive citance of the cited paper. This is set to 0 in case of a single citance.
17. **In how many different sections does the citation appear in the citing paper? (F17):** We take the number of different sections in which the citation to cited paper occurs and normalize it with the total number of sections present in the citing paper. Intuition is that if a citation occurs in most sections, it might be a significant citation.
18. **Number of common references in citing & cited paper normalized by the total number of references in citing article (F18):** We count the number of common bibliographic items present in the citing & cited paper and normalize it with total bibliographic items present in the citing paper.
19. **Number of common keywords between abstracts of the cited and citing paper extracted by YAKE [258] (F19):** We compare the number of common keyword between abstract of citing & cited paper extracted using YAKE. Our instinct is that more number of common keywords would denote more similarity between abstracts.
20. **Number of common keywords between titles of the cited and citing paper extracted by YAKE (F20):** We compare the number of common keywords between the title of citing & cited paper extracted using YAKE.
21. **Number of common keywords between the body of the cited and citing paper extracted by YAKE (F21):** We compare the number of common keyword between the body of citing & cited paper extracted using YAKE.
22. **Word Mover’s Distance (WMD) [259] between abstracts of the cited and citing paper (F22):** We measure the WMD between abstracts of citing & cited paper. The essence of this feature is to calculate semantic distance/similarity between abstracts of the two papers.
23. **WMD between titles of the cited and citing paper (F23):** We measure the WMD between title of citing & cited paper.

24. **WMD between the body of the cited and citing paper (F24):** We measure the WMD between the body of citing & cited paper.
25. **Average WMD between citance and abstract of the cited and citing paper (F25):** We take the average of WMDs between citance and abstract of the cited paper.
26. **Maximum WMD between citance and abstract of the cited and citing paper (F26):** We take the maximum of WMDs between citance and abstract of the cited paper.
27. **Average VADER [260] Polarity Index - Positive (F27), Negative (F28), Neutral (F29), Compound (F30):** We measure VADER polarity index of all the citance of cited paper, and take their average for each sentiment (positive, negative, neutral & compound).
28. **Maximum VADER Polarity Index - Positive (F31), Negative (F32), Neutral (F33), Compound (F34) of Citances:** We measure VADER polarity index of all the citance of cited paper, and take maximum among them for each sentiment (positive, negative, neutral & compound). The intuition to use sentiment information is to understand how the citing paper cites the cited paper.
29. **Number of common venues in Bibliography of citing and cited paper (F35):** We count the number of common venues mentioned in the bibliography of citing & cited paper and normalize it with the number of unique venues in citing paper. Higher venue overlap would signify that the papers are in the same domain [68].
30. **Number of common Authors in Bibliography of citing and cited paper (F36):** We count the number of common authors mentioned in the bibliography of citing & cited paper and normalize it with the number of unique authors in citing paper [68].

As mentioned earlier, only 14.3% of total citations are labeled as significant, which poses a *Class Imbalance* problem. To address this issue, we use SMOTE [261] along with random under-sampling of majority (contextual citation) class. We first split the dataset into 60% training & 40% testing data. Then we under-sample the majority class by 50%, and then we over-sample the minority class by 40%, on the training partition of the dataset.

#### 5.4.4 Evaluation

Our evaluation consists of two stages: first, we evaluate our approach on the citation significance task. Next, we try to see if we can identify the research lineage via tracing significant citations across the two research topics (*Document-level Novelty* and *MENNDL*). We ask the original authors of MENNDL to annotate the lineage and verify it with our automatic method. For

## 5.4 Finding a Research Lineage

document-novelty, we annotated the research lineage of our papers from Chapter 3. We train our model on the Valenzuela dataset and use that trained model to predict significant citations of *Document-level Novelty* and *MENNDL* papers, thereby try to visualize the research lineage across the citing papers. We curate a small citation graph to demonstrate our idea. Please note that our task in concern is Citation Significance Detection, which is different from Citation Classification in literature. Whereas Citation Classification focuses on to identify the intent of the citation, Citation Significance aims to identify the value associated with the citation. Obviously, the two tasks are related to each other, but the objectives are different.

### Citation Significance Detection

The goal of this task is to identify whether a citation was SIGNIFICANT or CONTEXTUAL. We experiment with several classifiers for the binary classification task ranging from kNN, Support Vector Machines, to Decision Trees. We found Random Forest to be the best performing one with our feature set. Table 5.10 shows our current results against the earlier reported results on the Valenzuela dataset. We attain promising results compared to earlier approaches with a relative improvement of 20 points in precision. Since the dataset is small, none of the earlier approaches or we attempted a deep neural approach on this dataset.

Table 5.11 shows classification results of the various classifiers we experimented with. Clearly, our features are highly inter-dependent, and hence it explains the better performance of Random Forests.

Table 5.10: Results on Citation Significance Detection on Valenzuela dataset

Methods	Precision
Valenzuela et al.[151]	0.65
Qayyum and Afzal et al.[148]	0.72
Nazir et al.[262]	0.75
Nazir et al.[263]	0.85
<b>Current Approach</b>	<b>0.92</b>

Table 5.11: Classification Result of various Classifiers for Citation Significance

Methods	Precision	Recall	F1-Score	Accuracy
kNN	0.80	0.87	0.83	0.81
SVM	0.79	0.67	0.73	0.81
Decision Tree	0.80	0.82	0.81	0.86
<b>Random Forest</b>	<b>0.92</b>	<b>0.82</b>	<b>0.87</b>	<b>0.90</b>

Figure 5.10 shows the importance of the top 10 features ranked as per their *information gain*.

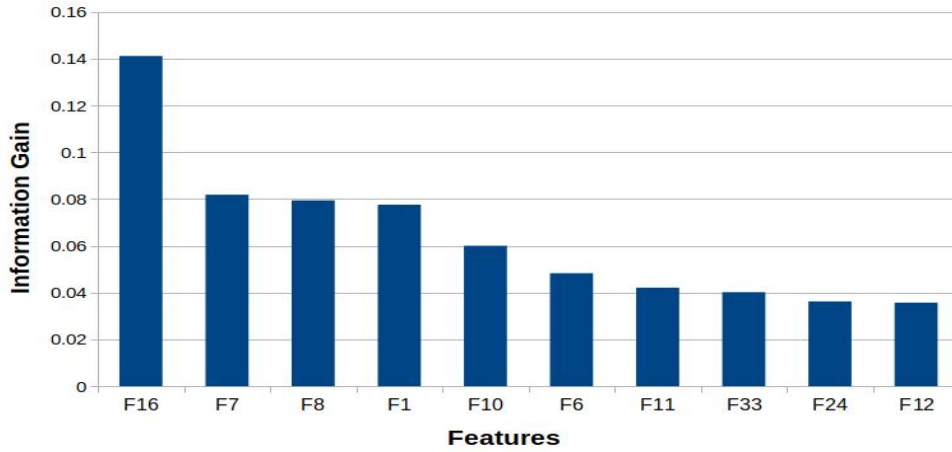


Figure 5.10: Feature importance ranked via Information Gain. Number of words between citances (F16) and Number of normalized citations (F7) are the most contributing features.

However, our experimental dataset is small, our features co-related, and hence it seems that some features have marginal contributions. We deem that in a real-life bigger dataset, the feature significance would be more visible. Here, we can see that features like *distance between citances*, *the number of concerned citation normalized by the total number of citations*, *similarity between cited-citing abstracts*, *in-text citation frequency*, *the similarity between citance & cited abstract*, play an important role in the classification. The other features featuring in the top 10 are: *number of citations from citing to cited normalized by the total citations made by the citing paper*, *max similarity between citance and cited abstract*, *neutral sentiment polarity of citances*, *the semantic distance between cited-citing pair (WMD)*, *similarity between citance and cited title*. We explain the possible reasons behind the performance of these features in Section 5.3.4. The precision with only using the top 10 features is 0.73. Hence, other features play a significant role, as well.

To mention here, authors in [264] found *Number of Direct Citations*, *Author Overlap*, and *Abstract Similarity* to be the most important features. Our approach performs good enough to proceed with the next stage.

### Research Lineage: Case Studies

Our end goal is not just citation classification but to make use of a highly accurate citation significance detection approach to trace significant citations and thereafter, try and establish a lineage of the given research. As explained earlier, by research lineage we aim to identify the idea propagation via tracking the significant citations. To achieve this, we create a *Significant Citation Graph*. A Significant Citation Graph (SCG) is a graph like structure, where each node

## 5.4 Finding a Research Lineage

---

represents a research paper and there is a directed edge between each cited-citing pair, whose direction is from cited paper node to citing paper node, indicating flow of knowledge from cited paper to citing paper. In an usual case, all citations have equal weights in a citation graph. However, in our case, each edge is labelled as either *significant* or *contextual*, using the approach we discussed in the previous section. Our idea is similar to that of existing scholarly graph databases, however, we go one step further and depict how a particular concept or *knowledge* has propagated with successive citations.

---

**Algorithm 4** Algorithm to Create Significance Citation Graph

---

**Input:** Trained Model & Concerned Research Paper  $P$

**Output:** Adjacency List for Citation Graph

Initialize adjacency list,  $A$

Initialize an empty queue,  $Q$

$Q.add(P)$

**while**  $Q$  is not empty

**for** each citation,  $C$  in  $Q[0]$  **do**

Extract features (F1-F36) for  $C$

**if**  $C$  is Significant **and**  $C$  is not in  $Q$  **then**

$Q.add(C)$

$A[Q[0]].add(C)$

$Q.pop()$

**return**  $A$

---

Algorithm 4 shows the method to create the adjacency list for the SCG. The Citation Significance Detection ML model is trained on a given dataset (Valenzuela in our case). To demonstrate the effectiveness of our method, we present a SCG for a set of papers on *Document-Level Novelty Detection* and *MENNDL*.

### Case Study I: Document-Level Novelty Detection

Figure 5.11 denotes an excerpt of a SCG from *Document-Level Novelty Detection* papers. The red edges denote *significant* citations whereas black edges denote *contextual* citations. Our approach determined if a citation edge is *significant* or *contextual*. In the citation graph, we are interested in the lineage among four textual novelty detection papers (P1, P2, P3, P4) which are annotated by the original authors (eventually us). P1 is the *pivot paper* where we introduce our document-level novelty detection dataset (TAP-DLND 1.0) and the other papers P2, P3, P4 are based on P1. While P2 and P4 address novelty classification, P3 aims to quantify textual novelty.

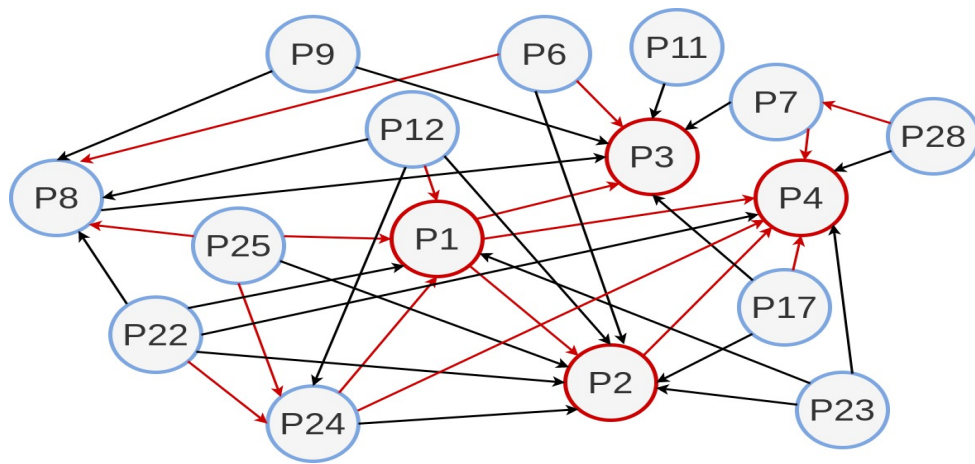


Figure 5.11: Significant Citation graph for a set of papers on *Document-Level Novelty Detection*. Please refer to the bibliography for the paper details. P1→[8], P2→[9], P3→[10], P4→[11], P6→[12], P7→[13], P8→[14], P9→[15], P11→[16], P12→[17], P17→[18], P22→[19], P23→[20], P24→[4], P25→[3], P28→[21]

Our current approach conforms to our annotations. With P1 as the pivot we can see that there are significant edges from P1 to each of P2, P3, and P4. There is also a significant edge between P2 and P4. However, there is no edge between P2 and P3, as they were contemporaneous submissions and their objective was different (P2 was about *novelty classification* and P3 was about *novelty scoring*).  $P1 \rightarrow P2 \rightarrow P4$  forms a research lineage as P2 extends on P1 and P4 extends on P2. Furthermore, we see that P12, P25, P24, P22 (transitively) are some influential papers for P1. P25 was the paper to introduce the first *document-level novelty detection* dataset but from an information retrieval perspective. P25 inspired us to create the dataset in P1 for ML experiments. We construe that P12, P22, P24 had significant influence for our investigations with P1. Hence, the current approach (trained on a different set of papers in Valenzuela dataset) proves successful to identify the significant citations and thereby also identify the corresponding lineage.

## Case Study II: MENNDL HPC Algorithm

We went ahead to test the efficacy of our approach to predict the *lineage* of a high-performance computing algorithm MENNDL. We show the research lineage of MENNDL [27] in Figure 5.12. We ask the original authors to annotate their research progression with MENNDL. As per the authors, the first paper to describe the MENNDL algorithm came in 2015 which is deemed as the *pivot* (P9). The follow-up paper that carried forward the work in P9 was P4 in 2017. Then P1 came in 2018 that built upon P4. P7, P12 came as extensions over P4. Next P6 came in 2019 that took forward the work from P1. With P9 as the source, our approach correctly predicted



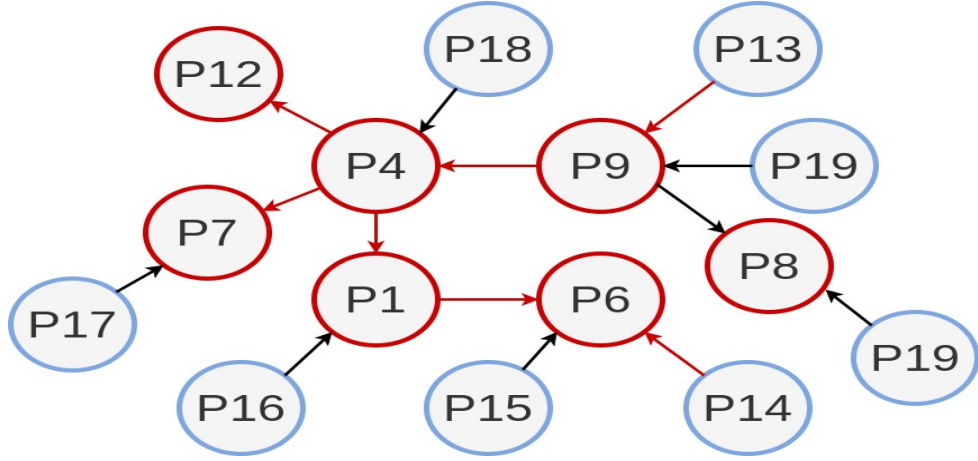


Figure 5.12: Significant Citation graph for a set of papers on *MENNDL HPC algorithm*. Please refer to the bibliography for the corresponding paper details.  $P1 \rightarrow [22]$ ,  $P4 \rightarrow [23]$ ,  $P6 \rightarrow [24]$ ,  $P7 \rightarrow [25]$ ,  $P8 \rightarrow [26]$ ,  $P9 \rightarrow [27]$ ,  $P12 \rightarrow [28]$ ,  $P13 \rightarrow [29]$ ,  $P14 \rightarrow [30]$ ,  $P15 \rightarrow [31]$ ,  $P16 \rightarrow [32]$ ,  $P17 \rightarrow [33]$ ,  $P18 \rightarrow [34]$ ,  $P19 \rightarrow [35]$ ,  $P20 \rightarrow [36]$ ,  $P25 \rightarrow [3]$ ,  $P28 \rightarrow [21]$

the lineage as  $P9 \rightarrow P4 \rightarrow P1 \rightarrow P6$ . Also, the lineage  $P9 \rightarrow P4 \rightarrow P12$  and  $P9 \rightarrow P4 \rightarrow P7$  via tracing significant citations could be clearly visible in the SCG at Figure 5.12. The authors annotated P8 as an application of P9 hence no significant link exists between P9 and P8.

From the above experiments and case studies, it is clear that our proposed method works reasonably well when a paper cites the influencing paper meaningfully. However, there are cases where some papers do not cite the papers from whom they are inspired. In such cases, our method would not work.

### 5.4.5 Conclusion and Future Work

Here, in this work we present our novel idea towards finding a research lineage to accelerate literature review. We achieve *state-of-the-art* performance on citation significance detection which is a crucial component to form the lineage. We leverage on that and show the efficacy of our approach on two completely different research topics. Our approach is simple and could be easily implemented on a large-scale citation graph (given the paper full-text). The training dataset is on NLP papers. However, we demonstrate the efficacy of our approach by testing on two topics: one from NLP and the other from HPC, hence establishing that our approach is domain-agnostic. Identifying significant citations to form a research lineage would also help the community to understand the true impact of a research beyond simple citation counts. We would look forward to experimenting with deep neural architectures to automatically identify meaningful features for the current task. Our next foray would be to identify the *missing citations* for papers which

may have played instrumental role in certain papers but unfortunately are not cited. This is particularly important in peer review to identify missing citations or unacknowledged work.

## 5.5 Limitations of our work

The work that we present in the current chapter are still mostly in the exploratory phase and are novel in their own right. As we can see that predicting the *fate of a paper* in peer review automatically is a very far-stretched vision. There are a multitude of factors, the most important being the human scientific knowledge which plays the central role in such decisions. This super important factor is missing from our approach and hence it would not be feasible to trust the predictions of such a system which base its decision on past paper decisions. Having said that, reviewer's attitude towards the paper can be ascertained from their *sentiment* reflected in their reviews. This problem opens the much larger problem: *how to evaluate the reviews?* Trusting the signals of a bad peer review would result in an arbitrary decision. Hence, the quality of a review should matter in the final decision on the paper just as the area chairs, program chairs, editors do; putting more weights into detailed, insightful reviews. Our proposed model is a very simple manifestation of the above process. However, we lack on annotations for peer review quality. Our next step would be to doing a minimal annotations on our data from experienced reviewers and bootstrapping from those supervisions. Our current evaluations are done qualitatively and would benefit from a quantitative evaluation. Also, we would envisage to put more thoughts on the *scoring* mechanism which is very simple as of now and may not truly reflect the exhaustiveness criteria we mention. The solution we propose for the final problem in this chapter would work well in case of papers where the significant citations are explicit or there is an evidence that one paper took forward the work of the other. Since the datasets are small we could not investigate automatic feature extraction via deep networks which probably would be a plausible direction to pursue next. Our current effort could only be seen as a *proof-of-concept* for establishing a *research lineage* via mining citation contexts from paper full-texts. There would be unforeseen challenges while we would port this idea of finding lineage via tracing significant citations to an actual large citation network. Information extraction from PDFs of scientific articles is still the first hurdle here. Another direction that is not addressed with the current method is to predict the *citation worthiness* of the scientific texts and recommend the missing citations. Without having the confidence that all required citations are correctly made from the paper, it would be difficult to establish a lineage that would be a true representative of influence propagation.

## 5.6 Chapter Summary

In this chapter we discuss three important problems related to AI in Peer Review. The first problem entails our venture to see how an AI trained on paper full-texts and reviews would perform to predict the peer-review outcome. With the second problem we delve into a more critical investigation to judge if a peer-review was significant enough for decision-making. Finally with the third problem we explore if we can find the lineage of a given research to assist reviewers to identify relevant prior art to judge the merit of the current paper under consideration.



## Conclusions and Future Works

### 6.1 Summary

Human scientific knowledge is documented in research papers and scientific reports. Peer Review is one such system that validates human scientific progress. However, with the exponential rise in research article submissions, present-day scholarly communications face several problems due to this information deluge. With the dawn of the age of Artificial Intelligence (AI), the world is intrigued if AI could move beyond the role of a simple digital assistant and attempt to understand complex human behavior and intellect. Peer review is one such AI-frontier task that demands a strong AI to understand the highest form of human intelligence manifested in research papers. The current research is dedicated to investigating some pressing problems relevant to peer review and scholarly communications. We summarize our investigations as:

- Chapter 3: *How much newness/novelty is there in a certain document? Can Natural Language Processing and Machine Learning help in identifying textual novelty?* We develop a document-level dataset for the task and discuss several approaches to detect the novelty of a document automatically.
- Chapter 4: *Can we determine with a reasonable degree of certainty if a research article falls within the scope of a journal? A considerable amount of resources (time, intellectual-labor) are wasted to reject a substantial number of papers sent to the wrong venues. Can an AI come to the rescue and thus speed up the overall peer-review process?* We develop several

Chapter	Problem	Contributions
Chapter 3	Textual Novelty Detection	TAP-DLND 1.0 dataset TAP-DLND 2.0 dataset Traditional feature-based model Deep Neural RDV-CNN model Decomposable attention-based model Novelty Scoring: SETDV-CNN Multi-premise entailment-based novelty detection
Chapter 3	Scope Detection	Feature-engineering method Deep multimodal neural architecture Multiobjective multiview clustering model
Chapter 3	AI for Peer Review	Deepsentipeer: Predicting peer review outcome A model to determine <i>peer review quality</i> Finding <i>research lineage</i> to accelerate relevant literature discovery

Table 6.1: Summarizing contribution of this thesis

approaches exploiting all available channels of information in a research paper to know its scope and thereby assist editors in locating *out-of-scope* submissions.

- Chapter 5: *Can we move beyond the quantitative measures of research quality (h-index/citation counts) and develop qualitative means to measure true research impact and pervasiveness in the community? Can that lead us to find the lineage of a given research?* We experiment with several features from the cited and the citing paper full-text to classify the nature of citations (significant or contextual). Leveraging this information, we demonstrate to form the research lineage for two different topics.
- Chapter 5: *Can an AI act as an editor and cumulate review information from human reviewers and generate a logical decision? Can we automatically estimate the quality of the peer reviews?* We explore if a deep neural model can extract features from paper full-text and reviews to predict the outcome of the peer-review process. We further develop an approach to see how exhaustive a review was and thereby judge its significance in decision-making.

## 6.2 Contributions of the Thesis

To sum up, we list our chapter-wise contributions in Table 6.1. Our contribution chapters are 3, 4, and 5.

## 6.3 Limitations of the Thesis

We have discussed on the limitations of this thesis in the respective chapters. To sum up, the shortcomings of the current thesis are as follows:

1. The novelty detection algorithms are not scalable for very large source document collections as the algorithms rely on  $n \times n$  comparisons with the source document sentences.
2. The novelty detection algorithms could not adapted in a straightforward manner for ascertaining scientific document novelty.
3. The role of image modality in scope detection is not very evident with the current dataset. Further experiments on image-rich papers like medicine, biology, astrophysics, etc. are required.
4. For automating/semi-automating decisions in the peer-review process, it is important to build *trust* on the peer reviews. Our current mechanism for deriving the *review quality score* is very simple. More investigations are required to make the scores representative of *review exhaustiveness* measure. Probably some annotations on the reviews from senior researchers in the field would help.
5. The idea of building a *research lineage* via tracing *significant citations* is elegant but would work on only a certain category of papers. Probably a more important problem is to investigate that whether a scientific statement missed a crucial citation which is quite common in scholarly communication. Hence, identifying missing scientific attribution appears to be an important direction which is currently not covered in this thesis.

## 6.4 Conclusions

Through chapters 3 to 5 we report our findings on the above problems. We enlist our core findings as:

- Identifying multi-premise entailment relationships between source and target texts is important for textual novelty detection.
- We can identify *out-of-scope* manuscripts leveraging on full-text and bibliographic information for certain journals to speed-up the peer-review process.
- We suggest that identifying the significance of the peer reviews is essential for making the final-decision (*not all reviews are equal*). Our investigation is a step towards that.

- We show that we can identify the lineage of the corresponding research for a certain category of papers via tracing significant citations. Our method could assist in relevant literature-discovery in the sea of papers.

The current research in no way supports the notion of an automated peer review system and strongly believes in human intellect to be in the loop. Despite the promise of AI overcoming or minimizing failings in the peer review process, AI will never entirely replace the human element. There are some things humans do better — or can at least oversee. Machines can and do make mistakes and malfunction. A human should be involved at various stages to prevent poor research from being accidentally published or good research from being wrongly denied publication. As with any computer involvement, hacking is an ever-present threat and can compromise the review process. Authors may change their writing style. Authors may learn how AI reviews research and change their writing to meet what they believe the algorithms are looking for. An AI algorithm will weigh scoring criteria but will need to determine which criteria it should weigh more heavily. AI is suitable for most quantitative analyses, but to analyze something qualitatively is more problematic. The quality of an article depends on more than just the validity and analysis of the data. It also depends on significance to existing research, innovation, impact on the field, writing style, and other nuances not easy for a machine to discern. AI can improve upon certain aspects of traditional peer review, and publishers are already using it for some basic tasks within the process. It will require definite policies and protocols for choosing which aspects of the review process can and should be automated, and when they need to apply human guidance or oversight. As technology continues to mature, more and more parts of the peer review process will likely benefit. However, at least for the foreseeable future, there will always be a need for human input and final decision-making.

A layman’s version of our research story could be found in the AWSAR-DST award-winning article: [https://www.awsar-dst.in/assets/winner\\_article\\_2018/43\\_PhD.pdf](https://www.awsar-dst.in/assets/winner_article_2018/43_PhD.pdf)

## 6.5 Future Work

Leveraging on the knowledge from this thesis, we would like to concentrate on the following as our future works:

- Creating a gold-standard dataset of research papers on a particular topic for scientific novelty detection. Automatically identifying premises of a given scientific claim and verifying if the claim is novel or not. It would be interesting to see the performance of our multi-premise novelty detection architecture on such data. The *gold standard dataset* for scientific novelty would involve subject-matter experts (researchers) in a certain area to



## 6.5 Future Work

---

annotate scientific papers for their *newness*. It would also involve creation of a knowledge base of articles that would serve as the background/source knowledge for that area. One important aspect here would be to identify articles that have no new knowledge, hence *non-novel*, and to identify their corresponding source articles. Novelty for scientific articles would not be a straightforward classification problem, probably it would make sense to determine the *degree of novelty* as novelty in research papers would significantly vary in proportion.

- Extending our work on peer review quality estimation with gold-standard annotations on actual peer review texts. Some experts would need to annotate the reviews for their perceived quality across several parameters.
- Integrating the investigation on research lineage to a large scholarly network to understand the academic impact and knowledge-flow for a given research topic. Our initial plan is to mine a subset of NLP papers from ACL Anthology pertaining to a certain area (e.g., sentiment analysis), create the corresponding citation network, and trace the *lineage* of papers within the network.



## CHAPTER 7

---

# Appendix

Sample documents are from the TAP-DLND 1.0 dataset. The dataset is available at <http://www.iitp.ac.in/~ai-nlp-ml/resources.html>

### Instance #45

#### Concatenated Source: 3 source documents concatenated to one

- 0: Raebareli (Uttar Pradesh) [India], November 1 (ANI): At least ten people were killed and 70 others sustained burn injuries after an ash-pipe exploded due to pressure at National Thermal Power Corporation (NTPC) plant in Unchahar area of Uttar Pradesh's Raebareli district.
- 1: Speaking to ANI, Additional Director General (Law and Order) Anand Kumar said, "As of now ten deaths have been confirmed by the district administration, while about 70 people have sustained burn injuries."
- 2: Following the incident, Uttar Pradesh Chief Minister Yogi Adityanath has announced an ex-gratia of Rs 2 lakh for the kin of deceased, Rs 50,000 for critically injured and Rs 25,000 for injured.
- 3: A 32-member team of National Disaster Response Force (NDRF) has left for Unchahar in Raebareli.
- 4: The authorities of NTPC said, "Rescue operations are underway in close coordination with District Administration.
- 5: Injured people have been shifted to nearby district hospitals.
- 6: An unfortunate accident in the boiler of 500 MW under trial unit of NTPC in Unchahar occurred this afternoon."
- 7: "NTPC's senior management is rushing to the site to coordinate the efforts," NTPC Corporation Communication Department said.
- 8: Union Health Minister J P Nadda spoke to UP Health Minister and Union Health Secretary to extend all possible help.
- 9: The blast reportedly took place when a boiler tube exploded at the unit.
- 10: "Ash-pipe exploded due to pressure at NTPC plant in Rae Bareli," the District Magistrate informed.
- 11: NTPC operates a 1550 Mega Watt power plant of Uttar Pradesh, which is named after Feroze Gandhi, the husband of former Prime Minister Indira Gandhi.
- 13: Englishmate Rahul Gandhi demands judicial probe into NTPC blast, meets victims 29 people died and over 60 injured in the boiler blast at the state-run power giant NTPCs Unchahar plant india Updated: Nov 02, 2017 19:15 IST Kenneth John Hindustan Times, Rae Bareli (Unchahar) Congress vice-president Rahul Gandhi arrives to meet the family members of the victims of Unchahar NTPC boiler blast in Raebareli on Thursday.
- 14: (PTI Photo ) Congress vice president Rahul Gandhi on Thursday demanded a judicial probe into the boiler blast at the state-run power giant NTPCs Unchahar plant in which 29 people died and over 60 injured.

- 15: He made the demand after visiting the district hospital in Rae Bareli, where the injured were undergoing treatment.
- 16: The accident showed serious lapse in working of the unit and to know the reason behind it a judicial probe was needed, Gandhi said during a brief interaction with media at the hospital.
- 17: Congress vice-president rushed to Rae Bareli taking a break from his ongoing Navsarjan Yatra in poll-bound Gujarat.
- 18: Rae Bareli is the parliamentary constituency of his mother and Congress president Sonia Gandhi.
- 19: To ascertain the cause of the blast, the NTPC has initiated a probe amid allegations from labourers that they had warned about the possible disaster at the ill-fated unit-six as the temperature near the furnace was steadily rising.
- 20: Gandhi consoled family members of victims who lost their lives in the blast and enquired about the condition of those admitted in the hospital.
- 21: Later, the Amethi member of parliament also visited private hospitals, SIMHANS and Nirmal, and enquired about condition of blast victims admitted there.
- 22: He also visited the site of boiler blast on NTPC premises and enquired about the accident from officials.
- 23: Senior Congress leader Ghulam Nabi Azad and UP Congress chief Raj Babbar accompanied the Congress vice president.
- 24: The laxities in construction should be probed thoroughly as without a major lapse accident of such magnitude was not possible, said Babbar.
- 25: The 500 megawatt unit 6 of the power plant was commissioned in April but due to technical fault in the boiler it failed to produce power.
- 26: We demand a judicial probe into the tragedy.
- 27: Setting up of an enquiry committee by government is just eyewash, reiterated Azad.
- 28: During his visit to the blast site Gandhi came face to face with union power minister RK Singh, who also visited the plant and took stock of the situation.
- 29: Singh denied any human negligence led to the blast.
- 30: Army pays tribute to 2 soldiers killed in Pulwama encounter Nov 03, 2017 16:55 IST .
- 31: NTPC shuts Unchahar plant unit after 26 die in blast; warnings ignored?
- 32: No human negligence behind boiler blast: Power Minister Agency Report — New Delhi/Rae Bareli — 2 November, 2017 — 11:30 PM Power producer NTPC has shut down a 500 MW unit at its Feroze Gandhi Unchahar Thermal Power Station in Rae Bareli in Uttar Pradesh following the accident on Wednesday that claimed the lives of 30 people.
- 33: Share this: Print The company in a regulatory filing on the BSE said: This is to inform that Unit 6 (500 MW) of Feroze Gandhi Unchahar Thermal Power Station, Rae Bareli, is under shut down after an accident in the evening of November 1, 2017.
- 34: The other five units of the station are operating normally. The death toll in the NTPC boiler blast here in Uttar Pradesh rose to 26 on Thursday, with more injured workers succumbing to their burn injuries, officials said.
- 35: The massive explosion took place in a 500 MW boiler unit in Unchahar town on the Lucknow-Allahabad highway.
- 36: Many were trapped when a fire erupted in the boiler and a huge ball of dust rose after the blast, making the rescue operations difficult.
- 37: On Thursday, contractual labourers at the plant raised slogans against the NTPC management.
- 38: They claimed they had warned about an impending disaster at unit six as the temperature near the furnace had been steadily rising.
- 39: The NTPC has launched a probe into the incident, which it said took place due to excess ash deposition in the furnace.
- 40: The state government has ordered a magisterial probe.
- 41: Union Power Minister R.K. Singh on Thursday denied claims by some political leaders and families of the deceased that human negligence was to blame for the boiler blast in NTPCs Unchahar unit here that left 30 dead and dozens seriously injured.
- 42: I have seen everything during my physical inspection of the accident scene and I can say that there is no human negligence in the unfortunate incident, Singh told reporters, after visiting the accident site along with state Power Minister Shrikant Sharma.
- 43: He also announced that the Central government has decided to give financial assistance of Rs 20 lakh to the families of the deceased and Rs 10 lakh each to the critically injured, while those who sustained minor injuries would get Rs 2 lakh each, the Union Minister announced.
- 44: This compensation would be in addition to the ex-gratia and financial assistance announced by Uttar Pradesh Chief

---

Minister Yogi Adityanath.

45: The state-run National Thermal Power Corp (NTPC) has also announced a financial assistance of Rs 5 lakh each to the families of the dead.

46: The Prime Ministers Relief Fund will also give Rs 2 lakh each to the next of kin of the deceased.

47: R.K. Singh also said that the priority of the government, both at the Centre and the state, was to save as many lives as possible, provide the best, prompt and adequate treatment.

48: Singh also said that NTPCs Unchahar unit was among the best in the country and that rumours that there was an extra load on it or that it was under pressure to increase production were unfounded and baseless.

49: How and why the accident happened would be conclusively found and detailed in the probe ordered by the Ministry which would be completed in 30 days, he added.

50: State Deputy Chief Minister Dinesh Sharma, who also visited Rae Bareilly on Thursday, urged the opposition parties not to make political currency out of the tragic incident.

51: Both the state and Union governments are saddened by the tragedy and are doing all they can to bring relief to the affected, he said and added that Prime Minister Narendra Modi was personally very sad at the loss of lives in the accident.

52: (IANS) Share this:

#### **Target Document**

0: The death toll from a blast at a coal-fired power plant in northern India rose to at least 29 on Thursday, as authorities launched an investigation into the cause of one of the countrys worst industrial accidents in years.

1: More than 20 survivors were battling for their lives with severe burns following Wednesdays blast in a newly-operated unit at the 1,550-megawatt plant run by state-owned NTPC, officials in Uttar Pradesh state said.

2: More than 80 others suffered injuries in the explosion.

3: Arvind Kumar, a principal secretary, said some of the severely injured had been taken to a hospital in the state capital Lucknow.

4: Blockages in the flue gas pipe in a unit led to the blast.

5: Hot flue gases and steam let out by the blast severely injured several workers

6: Sanjay Kumar Khatri, the top government official of Rae Bareilly district where the plant is located, told Reuters on Thursday.

7: A magisterial inquiry has been initiated.

8: This two-member technical team will submit findings within seven days, Khatri said.

9: In a statement, the National Human Rights Commission said an investigation was needed to ascertain whether negligence or errors had caused the explosion, and asked the state government to submit a detailed report within six weeks.

10: NTPC is the countrys top power producer and accidents have been rare at its facilities.

11: Senior state police official Anand Kumar said on Wednesday ash had piled up in the furnace beneath the boiler, which led to a build-up of pressure resulting in the explosion.

12: The power ministry and state government have both offered cash compensation to the families of the deceased and to the injured.

13: The plant in the town of Unchahar supplies electricity to nine states, NTPCs website showed.

14: The company said other facilities would make up for the shortfall and outages were unlikely.

15: The 500 MW unit had been operating since April and was shut down after the accident.

16: The other five units of the station are operating normally, NTPC said in a statement.

17: NTPC has initiated an inquiry into the incident.

18: We are not a company that will take any risk.

19: We have so many units that if power cannot be supplied by one, it can be given by the other.

20: It was a sudden accident, an NTPC official, who did not wish to be named, said.

My Splitter

Upload the .txt file Open the sources Edit Target file

Sentence	Feedback
` Tragedy King ' Dilip Kumar admitted to Mumbai 's Lilavati hospital The ` Naya Daur ' star actor has been facing medical complications in recent years August 2 , 2017 Last Updated at 23:33 IST email this article Type address separated by commas Your Email : Enter the characters shown in the image .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
Send me a copy : Dilip Kumar ( Photo Credit : Filmfare ) ALSO READ Veteran Bollywood actor Dilip Kumar has been admitted to Mumbai 's Lilavati hospital .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
Dr Jalil Parkar , who generally treats him , told ANI that the 94-year-old actor has been admitted to Lilavati hospital and tests are being conducted on him .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
The ` Naya Daur ' star actor has been facing medical complications in recent years .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
Previously in 2016 , he was hospitalised in April due to fever and nausea .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
Known as the ` Tragedy King ' , Kumar has acted in over 65 films in his career .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
Spanning a career of over six decades , the ` Kranti ' star has done almost 65 films .	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED
( Only the headline and picture of this report may have been reworked by the Business Standard staff ; the rest of the content is auto-generated from a syndicated feed . )	<input type="radio"/> IRR <input type="radio"/> NOV <input type="radio"/> NN <input type="radio"/> PN25 <input type="radio"/> PN50 <input type="radio"/> PN75 <input checked="" type="radio"/> UNMARKED

Submit

Figure 7.1: TAP-DLND 2.0 Annotation Interface

# References

---

- [1] J. L. Fleiss, “Measuring nominal scale agreement among many raters.” *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.
- [2] T. Ghosal, A. Salam, S. Tiwary, A. Ekbal, and P. Bhattacharyya, “TAP-DLND 1.0 : A corpus for document level novelty detection,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, 2018.
- [3] Y. Zhang, J. Callan, and T. Minka, “Novelty and redundancy detection in adaptive filtering,” in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 81–88.
- [4] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis, “Efficient online novelty detection in news streams.” in *WISE (1)*, 2013, pp. 57–71.
- [5] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. H. Hovy, and R. Schwartz, “A dataset of peer reviews (peerread): Collection, insights and NLP applications,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, 2018, pp. 1647–1661. [Online]. Available: <https://aclanthology.info/papers/N18-1149/n18-1149>
- [6] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 2017, pp. 670–680. [Online]. Available: <https://aclanthology.info/papers/D17-1070/d17-1070>
- [7] C. J. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*, 2014. [Online]. Available: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109>
- [8] T. Ghosal, A. Salam, S. Tiwari, A. Ekbal, and P. Bhattacharyya, “Tap-dlnd 1.0: A corpus for document level novelty detection,” *arXiv preprint arXiv:1802.06950*, 2018.
- [9] T. Ghosal, V. Edithal, A. Ekbal, P. Bhattacharyya, G. Tsatsaronis, and S. S. S. K. Chivukula, “Novelty goes deep. a deep neural solution to document level novelty detection,” in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2802–2813.
- [10] T. Ghosal, A. Shukla, A. Ekbal, and P. Bhattacharyya, “To comprehend the new: On measuring the freshness of a document,” in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.
- [11] T. Ghosal, V. Edithal, A. Ekbal, P. Bhattacharyya, S. S. S. K. Chivukula, and G. Tsatsaronis, “Is your document novel? let attention guide you. an attention based model for document level novelty detection,” *Natural Language Engineering*, vol. 1, no. 1, pp. 1–38, 1997.
- [12] I. Soboroff and D. Harman, “Overview of the trec 2003 novelty track.” in *TREC*, 2003, pp. 38–53.
- [13] P. Zhao and D. L. Lee, “How much novelty is relevant? it depends on your curiosity,” in *39th International ACM SIGIR Conference on Research and Development, Pisa, Italy*, 2016, p. 100.
- [14] R. Colomo-Palacios, F. S. Tsai, and K. L. Chan, “Redundancy and novelty mining in the business blogosphere,” *The Learning Organization*, 2010.
- [15] W. Tang, F. S. Tsai, and L. Chen, “Blended metrics for novel sentence mining,” *Expert Systems with Applications*, vol. 37, no. 7, pp. 5172–5177, 2010.

- [16] M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger *et al.*, “From word embeddings to document distances,” in *ICML*, vol. 15, 2015, pp. 957–966.
- [17] X. Li and W. B. Croft, “Novelty detection based on sentence level patterns,” in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 744–751.
- [18] B. Schiffman and K. R. McKeown, “Context and learning in novelty detection,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 716–723.
- [19] J. Allan, C. Wade, and A. Bolivar, “Retrieval and novelty detection at the sentence level,” in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 314–321.
- [20] I. Soboroff and D. Harman, “Novelty detection: the trec experience,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 105–112.
- [21] F. Zhang, K. Zheng, N. J. Yuan, X. Xie, E. Chen, and X. Zhou, “A novelty-seeking based dining recommender system,” in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1362–1372.
- [22] R. M. Patton, J. T. Johnston, S. R. Young, C. D. Schuman, D. D. March, T. E. Potok, D. C. Rose, S.-H. Lim, T. P. Karnowski, M. A. Ziatdinov *et al.*, “167-pflops deep learning for electron microscopy: from learning physics to atomic manipulation,” in *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2018, pp. 638–648.
- [23] S. R. Young, D. C. Rose, T. Johnston, W. T. Heller, T. P. Karnowski, T. E. Potok, R. M. Patton, G. Perdue, and J. Miller, “Evolving deep networks using hpc,” in *Proceedings of the Machine Learning on HPC Environments*, 2017, pp. 1–7.
- [24] R. M. Patton, J. T. Johnston, S. R. Young, C. D. Schuman, T. E. Potok, D. C. Rose, S.-H. Lim, J. Chae, L. Hou, S. Abousamra *et al.*, “Exascale deep learning to accelerate cancer research,” in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 1488–1496.
- [25] J. T. Johnston, S. R. Young, C. D. Schuman, J. Chae, D. D. March, R. M. Patton, and T. E. Potok, “Fine-grained exploitation of mixed precision for faster cnn training,” in *2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)*. IEEE, 2019, pp. 9–18.
- [26] T. Johnston, S. R. Young, D. Hughes, R. M. Patton, and D. White, “Optimizing convolutional neural networks for cloud detection,” in *Proceedings of the Machine Learning on HPC Environments*, 2017, pp. 1–9.
- [27] S. R. Young, D. C. Rose, T. P. Karnowski, S.-H. Lim, and R. M. Patton, “Optimizing deep learning hyperparameters through an evolutionary algorithm,” in *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments*, 2015, pp. 1–5.
- [28] J. Chae, C. D. Schuman, S. R. Young, J. T. Johnston, D. C. Rose, R. M. Patton, and T. E. Potok, “Visualization system for evolutionary neural networks for deep learning,” in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 4498–4502.
- [29] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675–678.
- [30] J. Saltz, R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, D. Samaras, K. R. Shroyer, T. Zhao, R. Batiste *et al.*, “Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images,” *Cell reports*, vol. 23, no. 1, pp. 181–193, 2018.
- [31] V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. O. Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy *et al.*, “The immune landscape of cancer,” *Immunity*, vol. 48, no. 4, pp. 812–830, 2018.
- [32] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT’2010*. Springer, 2010, pp. 177–186.
- [33] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.
- [34] A. Lucchi, P. Márquez-Neila, C. Becker, Y. Li, K. Smith, G. Knott, and P. Fua, “Learning structured models for segmentation of 2-d and 3-d imagery,” *IEEE transactions on medical imaging*, vol. 34, no. 5, pp. 1096–1110, 2014.



- 
- [35] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, vol. 5, no. 1, pp. 1–9, 2014.
- [36] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3642–3649.
- [37] E. Landhuis, "Scientific literature: Information overload," *Nature*, vol. 535, no. 7612, pp. 457–458, 2016.
- [38] J. Beall, "Predatory publishing is just one of the consequences of gold open access," *Learned Publishing*, vol. 26, no. 2, pp. 79–84, 2013.
- [39] T. F. Frandsen, "Why do researchers decide to publish in questionable journals? a review of the literature," *Learned Publishing*, vol. 32, no. 1, pp. 57–62, 2019.
- [40] H. A. Maurer, F. Kappe, and B. Zaka, "Plagiarism-a survey." *J. UCS*, vol. 12, no. 8, pp. 1050–1084, 2006.
- [41] E. A. Fong and A. W. Wilhite, "Authorship and citation manipulation in academic research," *PLoS One*, vol. 12, no. 12, p. e0187394, 2017.
- [42] B.-C. Björk and D. Solomon, "The publishing delay in scholarly peer-reviewed journals," *Journal of informetrics*, vol. 7, no. 4, pp. 914–923, 2013.
- [43] M. L. Cooper, "Problems, pitfalls, and promise in the peer-review process: Commentary on trafimow & rice (2009)," *Perspectives on Psychological Science*, vol. 4, no. 1, pp. 84–90, 2009.
- [44] P. Cohen, "Scholars test web alternative to peer review," *The New York Times*, vol. 23, 2010.
- [45] C. Faggion, "Improving the peer-review process from the perspective of an author and reviewer," *British Dental Journal*, vol. 220, no. 4, pp. 167–168, 2016.
- [46] J. P. Ioannidis, "Meta-research: Why research on research matters," *PLoS biology*, vol. 16, no. 3, p. e2005468, 2018.
- [47] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi, "Ai-assisted peer review," *Humanities and Social Sciences Communications*, vol. 8, no. 1, pp. 1–11, 2021.
- [48] M. J. Mrowinski, P. Fronczak, A. Fronczak, M. Ausloos, and O. Nedic, "Artificial intelligence in peer review: How can evolutionary computation support journal editors?" *PloS one*, vol. 12, no. 9, p. e0184711, 2017.
- [49] L. Charlin and R. Zemel, "The toronto paper matching system: an automated paper-reviewer assignment system," 2013.
- [50] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 500–509.
- [51] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin, "Bias in peer review," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 1, pp. 2–17, 2013.
- [52] T. Ross-Hellauer, "What is open peer review? a systematic review," *F1000Research*, vol. 6, 2017.
- [53] G. Helgesson and S. Eriksson, "Plagiarism in research," *Medicine, Health Care and Philosophy*, vol. 18, no. 1, pp. 91–101, 2015.
- [54] H. V. Halteren, "Author verification by linguistic profiling: An exploration of the parameter space," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 4, no. 1, pp. 1–17, 2007.
- [55] Y. Dong, R. A. Johnson, and N. V. Chawla, "Will this paper increase your h-index? scientific impact prediction," in *Proceedings of the eighth ACM international conference on web search and data mining*, 2015, pp. 149–158.
- [56] M. B. Nuijten and J. R. Polanin, "'statcheck': Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses," *Research Synthesis Methods*, 2020.
- [57] J. L. Cornelius, "Reviewing the review process: Identifying sources of delay," *The Australasian Medical Journal*, vol. 5, no. 1, p. 26, 2012.
- [58] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: learning to estimate future citations for literature," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1247–1252.
- [59] A. Bartoli, A. De Lorenzo, E. Medvet, and F. Tarlao, "Your paper has been accepted, rejected, or whatever: Automatic generation of scientific paper reviews," in *International Conference on Availability, Reliability, and Security*. Springer, 2016, pp. 19–28.
- [60] S. A. Greenberg, "How citation distortions create unfounded authority: analysis of a citation network," *Bmj*, vol. 339, p. b2680, 2009.

- 
- [61] P. Dall’Aglio, “Peer review and journal models,” *arXiv preprint physics/0608307*, 2006.
  - [62] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, “Mastering the game of go without human knowledge,” *nature*, vol. 550, no. 7676, pp. 354–359, 2017.
  - [63] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland *et al.*, “Improved protein structure prediction using potentials from deep learning,” *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.
  - [64] T. Ghosal, V. Edithal, A. Ekbal, P. Bhattacharyya, G. Tsatsaronis, and S. S. S. K. Chivukula, “Novelty goes deep. A deep neural solution to document level novelty detection,” in *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 2018, pp. 2802–2813. [Online]. Available: <https://aclanthology.info/papers/C18-1237/c18-1237>
  - [65] T. Ghosal, A. Shukla, A. Ekbal, and P. Bhattacharyya, “To comprehend the new: On measuring the freshness of a document,” in *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*. IEEE, 2019, pp. 1–8. [Online]. Available: <https://doi.org/10.1109/IJCNN.2019.8851857>
  - [66] T. Ghosal, V. Edithal, A. Ekbal, P. Bhattacharyya, S. S. S. K. Chivukula, and G. Tsatsaronis, “Is your document novel? let attention guide you. an attention-based model for document-level novelty detection,” *Natural Language Engineering*, p. 1–28, 2020.
  - [67] S. Trumbore, M.-E. Carr, and S. Mikaloff-Fletcher, “Criteria for rejection of papers without review,” *Global Biogeochemical Cycles*, vol. 29, no. 8, pp. 1123–1123, 2015.
  - [68] T. Ghosal, R. Sonam, A. Ekbal, S. Saha, and P. Bhattacharyya, “Is the paper within scope? are you fishing in the right pond?” in *19th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2019, Champaign, IL, USA, June 2-6, 2019*, M. Bonn, D. Wu, J. S. Downie, and A. Martaus, Eds. IEEE, 2019, pp. 237–240. [Online]. Available: <https://doi.org/10.1109/JCDL.2019.00040>
  - [69] T. Ghosal, R. Verma, A. Ekbal, S. Saha, and P. Bhattacharyya, “An empirical study of importance of different sections in research articles towards ascertaining their appropriateness to a journal,” in *Digital Libraries at Times of Massive Societal Transition*, E. Ishita, N. L. S. Pang, and L. Zhou, Eds. Cham: Springer International Publishing, 2020, pp. 407–415.
  - [70] T. Ghosal, A. Raj, A. Ekbal, S. Saha, and P. Bhattacharyya, “A deep multimodal investigation to determine the appropriateness of scholarly submissions,” in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2019, pp. 227–236.
  - [71] T. Ghosal, D. Dey, A. Dutta, A. Ekbal, S. Saha, and P. Bhattacharyya, “A multiview clustering approach to identify out-of-scope submissions in peer review,” in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2019, pp. 392–393.
  - [72] T. Ghosal, R. Verma, A. Ekbal, and P. Bhattacharyya, “Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1120–1130.
  - [73] C. L. Wayne, “Topic detection and tracking (tdt),” in *Workshop held at the University of Maryland on*, vol. 27. Citeseer, 1997, p. 28.
  - [74] J. Allan, R. Papka, and V. Lavrenko, “On-line new event detection and tracking,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 37–45.
  - [75] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, “Topic-conditioned novelty detection,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 688–693.
  - [76] N. Stokes and J. Carthy, “First story detection using a composite document representation,” in *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 2001, pp. 1–8.
  - [77] M. Franz, A. Ittycheriah, J. S. McCarley, and T. Ward, “First story detection: Combining similarity and novelty based approaches,” in *Topic Detection and Tracking Workshop Report*, 2001, pp. 193–206.
  - [78] J. Allan, V. Lavrenko, D. Malin, and R. Swan, “Detections, bounds, and timelines: Umass and tdt-3,” in *Proceedings of topic detection and tracking workshop*, 2000, pp. 167–174.
  - [79] Y. Yang, T. Pierce, and J. Carbonell, “A study of retrospective and on-line event detection,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 28–36.

- 
- [80] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 330–337.
  - [81] A. Bagga and B. Baldwin, "Cross-document event coreference: Annotations, experiments, and observations," in *Coreference and Its Applications*, 1999.
  - [82] D. Harman, "Overview of the trec 2002 novelty track." in *TREC*, 2002.
  - [83] I. Soboroff, "Overview of the TREC 2004 novelty track," in *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, 2004. [Online]. Available: <http://trec.nist.gov/pubs/trec13/papers/NOVELTY.OVERVIEW.pdf>
  - [84] A. T. Kwee, F. S. Tsai, and W. Tang, "Sentence-level novelty detection in english and malay," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2009, pp. 40–51.
  - [85] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, L. Zhao, and S. Ma, "Expansion-based technologies in finding relevant and new information: Thu trec 2002: Novelty track experiments," *NIST SPECIAL PUBLICATION SP*, no. 251, pp. 586–590, 2003.
  - [86] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan, "Information filtering, novelty detection, and named-page finding." in *TREC*, 2002.
  - [87] E. Gabrilovich, S. Dumais, and E. Horvitz, "Newsjunkie: providing personalized newsfeeds via analysis of information novelty," in *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004, pp. 482–490.
  - [88] Y. Zhang and F. S. Tsai, "Combining named entities and tags for novel sentence detection," in *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*. ACM, 2009, pp. 30–34.
  - [89] L. Ru, L. Zhao, M. Zhang, and S. Ma, "Improved feature selection and redundance computing-thuir at trec 2004 novelty track." in *TREC*, 2004.
  - [90] M. Gamon, "Graph-based text representation for novelty detection," in *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 17–24.
  - [91] F. S. Tsai, W. Tang, and K. L. Chan, "Evaluation of novelty metrics for sentence-level novelty mining," *Information Sciences*, vol. 180, no. 12, pp. 2359–2374, 2010.
  - [92] F. S. Tsai and K. Luk Chan, "Redundancy and novelty mining in the business blogosphere," *The Learning Organization*, vol. 17, no. 6, pp. 490–499, 2010.
  - [93] L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo, "The seventh pascal recognizing textual entailment challenge." in *TAC*, 2011.
  - [94] I. Dagan, D. Roth, M. Sammons, and F. M. Zanzotto, "Recognizing textual entailment: Models and applications," *Synthesis Lectures on Human Language Technologies*, vol. 6, no. 4, pp. 1–220, 2013.
  - [95] F. S. Tsai and Y. Zhang, "D2s: Document-to-sentence framework for novelty detection," *Knowledge and information systems*, vol. 29, no. 2, pp. 419–433, 2011.
  - [96] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis, "Efficient online novelty detection in news streams," in *Web Information Systems Engineering - WISE 2013 - 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part I*, 2013, pp. 57–71. [Online]. Available: [https://doi.org/10.1007/978-3-642-41230-1\\_5](https://doi.org/10.1007/978-3-642-41230-1_5)
  - [97] A. Verheij, A. Kleijn, F. Frasincar, and F. Hogenboom, "A comparison study for novelty control mechanisms applied to web news stories," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2012, pp. 431–436.
  - [98] T. Dasgupta and L. Dey, "Automatic scoring for innovativeness of textual ideas," in *Knowledge Extraction from Text, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 12, 2016.*, 2016. [Online]. Available: <http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12663>
  - [99] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 335–336.
  - [100] P. Chandar and B. Carterette, "Preference based evaluation measures for novelty and diversity," in *SIGIR*, 2013.

- [101] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *SIGIR*, 2008.
- [102] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, "A comparative analysis of cascade measures for novelty and diversity," in *WSDM*, 2011.
- [103] N. Kang, M. A. Doornenbal, and R. J. Schijvenaars, "Elsevier journal finder: recommending journals for your paper," in *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 2015, pp. 261–264.
- [104] M. Errami, J. D. Wren, J. M. Hicks, and H. R. Garner, "etblast: a web server to identify expert reviewers, appropriate journals and similar publications," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W12–W15, 2007.
- [105] A. Doms and M. Schroeder, "Gopubmed: exploring pubmed with the gene ontology," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W783–W786, 2005.
- [106] A. D. Eaton, "Hubmed: a web-based biomedical literature search interface," *Nucleic acids research*, vol. 34, no. suppl 2, pp. W745–W747, 2006.
- [107] T. Goetz and C.-W. von der Lieth, "Pubfinder: a tool for improving retrieval rate of relevant pubmed abstracts," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W774–W778, 2005.
- [108] H. Alhoori and R. Furuta, "Recommendation of scholarly venues based on dynamic user interests," *Journal of Informetrics*, vol. 11, no. 2, pp. 553–563, 2017.
- [109] I. Boukhris and R. Ayachi, "A novel personalized academic venue hybrid recommender," in *Computational Intelligence and Informatics (CINTI), 2014 IEEE 15th International Symposium on*. IEEE, 2014, pp. 465–470.
- [110] H. Luong, T. Huynh, S. Gauch, L. Do, and K. Hoang, "Publication venue recommendation using author network's publication history," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2012, pp. 426–435.
- [111] S. Yu, J. Liu, Z. Yang, Z. Chen, H. Jiang, A. Tolba, and F. Xia, "Pave: Personalized academic venue recommendation exploiting co-publication networks," *Journal of Network and Computer Applications*, vol. 104, pp. 38–47, 2018.
- [112] Z. Chen, F. Xia, H. Jiang, H. Liu, and J. Zhang, "Aver: Random walk based academic venue recommendation," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 579–584.
- [113] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 439–448.
- [114] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, pp. 104–116, 2015.
- [115] S. Price and P. A. Flach, "Computational support for academic peer review: a perspective from artificial intelligence," *Commun. ACM*, vol. 60, no. 3, pp. 70–79, 2017. [Online]. Available: <https://doi.org/10.1145/2979672>
- [116] K. Wang and X. Wan, "Sentiment analysis of peer review texts for scholarly papers," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, 2018, pp. 175–184. [Online]. Available: <https://doi.org/10.1145/3209978.3210056>
- [117] A. C. Justice, M. K. Cho, M. A. Winker, J. A. Berlin, D. Rennie, P. Investigators *et al.*, "Does masking author identity improve peer review quality?: A randomized controlled trial," *Jama*, vol. 280, no. 3, pp. 240–242, 1998.
- [118] S. Schroter, N. Black, S. Evans, J. Carpenter, F. Godlee, and R. Smith, "Effects of training on quality of peer review: randomised controlled trial," *Bmj*, vol. 328, no. 7441, p. 673, 2004.
- [119] T. Jefferson, E. Wager, and F. Davidoff, "Measuring the quality of editorial peer review," *Jama*, vol. 287, no. 21, pp. 2786–2790, 2002.
- [120] S. van Rooyen, N. Black, and F. Godlee, "Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts," *Journal of clinical epidemiology*, vol. 52, no. 7, pp. 625–629, 1999.
- [121] M. M. Shattell, P. Chinn, S. P. Thomas, and W. R. Cowling III, "Authors' and editors' perspectives on peer review quality in three scholarly nursing journals," *Journal of Nursing Scholarship*, vol. 42, no. 1, pp. 58–65, 2010.

- 
- [122] S. Van Rooyen, "The evaluation of peer-review quality," *Learned Publishing*, vol. 14, no. 2, pp. 85–91, 2001.
  - [123] D. Houry, S. Green, and M. Callaham, "Does mentoring new peer reviewers improve review quality? a randomized trial," *BMC Medical Education*, vol. 12, no. 1, p. 83, 2012.
  - [124] R. Bruce, A. Chauvin, L. Trinquart, P. Ravaud, and I. Boutron, "Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis," *BMC medicine*, vol. 14, no. 1, p. 85, 2016.
  - [125] M. Enserink, "Peer review and quality: A dubious connection?" 2001.
  - [126] R. D'Andrea and J. P. O'Dwyer, "Can editors save peer review from peer reviewers?" *PloS one*, vol. 12, no. 10, p. e0186111, 2017.
  - [127] D. Rennie, "Let's make peer review scientific," *Nature News*, vol. 535, no. 7610, p. 31, 2016.
  - [128] M. L. Callaham, W. G. Baxt, J. F. Waeckerle, and R. L. Wears, "Reliability of editors' subjective quality ratings of peer reviews of manuscripts," *Jama*, vol. 280, no. 3, pp. 229–231, 1998.
  - [129] A. Sizo, A. Lino, L. P. Reis, and Á. Rocha, "An overview of assessing the quality of peer review reports of scientific articles," *International Journal of Information Management*, vol. 46, pp. 286–293, 2019.
  - [130] D. Sculley, J. Snoek, and A. B. Wiltschko, "Avoiding a tragedy of the commons in the peer review process," *CoRR*, vol. abs/1901.06246, 2019. [Online]. Available: <http://arxiv.org/abs/1901.06246>
  - [131] C. Superchi, J. A. González, I. Solà, E. Cobo, D. Hren, and I. Boutron, "Tools used to assess the quality of peer review reports: a methodological systematic review," *BMC medical research methodology*, vol. 19, no. 1, p. 48, 2019.
  - [132] J. M. Wicherts, "Peer review quality and transparency of the peer-review process in open access and subscription journals," *PloS one*, vol. 11, no. 1, p. e0147913, 2016.
  - [133] S. Van Rooyen, F. Godlee, S. Evans, N. Black, and R. Smith, "Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial," *Bmj*, vol. 318, no. 7175, pp. 23–27, 1999.
  - [134] W. Xiong and D. Litman, "Automatically predicting peer-review helpfulness," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 502–507.
  - [135] X. Zhu, P. D. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 2, pp. 408–427, 2015. [Online]. Available: <https://doi.org/10.1002/asi.23179>
  - [136] C. Shi, H. Wang, B. Chen, Y. Liu, and Z. Zhou, "Visual analysis of citation context-based article influence ranking," *IEEE Access*, vol. 7, pp. 113 853–113 866, 2019. [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2932051>
  - [137] Y. Xie, Y. Sun, and L. Shen, "Predicating paper influence in academic network," in *20th IEEE International Conference on Computer Supported Cooperative Work in Design, CSCWD 2016, Nanchang, China, May 4-6, 2016*. IEEE, 2016, pp. 539–544. [Online]. Available: <https://doi.org/10.1109/CSCWD.2016.7566047>
  - [138] J. Shen, Z. Song, S. Li, Z. Tan, Y. Mao, L. Fu, L. Song, and X. Wang, "Modeling topic-level academic influence in scientific literatures," in *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 13, 2016*, ser. AAAI Workshops, M. Khabsa, C. L. Giles, and A. D. Wade, Eds., vol. WS-16-13. AAAI Press, 2016. [Online]. Available: <http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12598>
  - [139] G. Manju, V. Kavitha, and T. V. Geetha, "Influential researcher identification in academic network using rough set based selection of time-weighted academic and social network features," *Int. J. Intell. Inf. Technol.*, vol. 13, no. 1, pp. 1–25, 2017. [Online]. Available: <https://doi.org/10.4018/IJIT.2017010101>
  - [140] S. F. Pileggi, "Looking deeper into academic citations through network analysis: popularity, influence and impact," *Univers. Access Inf. Soc.*, vol. 17, no. 3, pp. 541–548, 2018. [Online]. Available: <https://doi.org/10.1007/s10209-017-0565-5>
  - [141] F. Zhang and S. Wu, "Predicting future influence of papers, researchers, and venues in a dynamic academic network," *J. Informetrics*, vol. 14, no. 2, p. 101035, 2020. [Online]. Available: <https://doi.org/10.1016/j.joi.2020.101035>
  - [142] C. Ji, Y. Tang, and G. Chen, "Analyzing the influence of academic papers based on improved pagerank," in *Emerging Technologies for Education - 4th International Symposium, SETE@ICWL 2019, Magdeburg, Germany, September 23-25, 2019, Revised Selected Papers*, ser. Lecture Notes in Computer Science, E. Popescu, T. Hao, T. Hsu, H. Xie, M. Temperini, and W. Chen, Eds., vol. 11984. Springer, 2019, pp. 214–225. [Online]. Available: [https://doi.org/10.1007/978-3-030-38778-5\\_24](https://doi.org/10.1007/978-3-030-38778-5_24)

- [143] F. Wang, C. Jia, J. Liu, and J. Liu, “Dynamic assessment of the academic influence of scientific literature from the perspective of altmetrics,” in *Proceedings of the 17th International Conference on Scientometrics and Informetrics, ISSI 2019, Rome, Italy, September 2-5, 2019*, G. Catalano, C. Daraio, M. Gregori, H. F. Moed, and G. Ruocco, Eds. ISSI Society, 2019, pp. 2528–2529.
- [144] F. Zhao, Y. Zhang, J. Lu, and O. Shai, “Measuring academic influence using heterogeneous author-citation networks,” *Scientometrics*, vol. 118, no. 3, pp. 1119–1140, 2019. [Online]. Available: <https://doi.org/10.1007/s11192-019-03010-5>
- [145] C. Dong and U. Schäfer, “Ensemble-style self-training on citation classification,” in *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*. The Association for Computer Linguistics, 2011, pp. 623–631. [Online]. Available: <https://www.aclweb.org/anthology/I11-1070/>
- [146] S. Teufel, A. Siddharthan, and D. Tidhar, “Automatic classification of citation function,” in *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, D. Jurafsky and É. Gaussier, Eds. ACL, 2006, pp. 103–110. [Online]. Available: <https://www.aclweb.org/anthology/W06-1613/>
- [147] M. H. Alvarez, J. M. G. Soriano, and P. Martínez-Barco, “Citation function, polarity and influence classification,” *Nat. Lang. Eng.*, vol. 23, no. 4, pp. 561–588, 2017. [Online]. Available: <https://doi.org/10.1017/S1351324916000346>
- [148] F. Qayyum and M. T. Afzal, “Identification of important citations by exploiting research articles’ metadata and cue-terms from content,” *Scientometrics*, vol. 118, no. 1, pp. 21–43, 2019. [Online]. Available: <https://doi.org/10.1007/s11192-018-2961-x>
- [149] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady, “Structural scaffolds for citation intent classification in scientific publications,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 3586–3596. [Online]. Available: <https://doi.org/10.18653/v1/n19-1361>
- [150] D. Pride and P. Knoth, “An authoritative approach to citation classification,” in *JCDL ’20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*, R. Huang, D. Wu, G. Marchionini, D. He, S. J. Cunningham, and P. Hansen, Eds. ACM, 2020, pp. 337–340. [Online]. Available: <https://doi.org/10.1145/3383583.3398617>
- [151] M. Valenzuela, V. Ha, and O. Etzioni, “Identifying meaningful citations,” in *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January, 2015*, ser. AAAI Workshops, C. Caragea, C. L. Giles, N. L. Bhamidipati, D. Caragea, S. D. Gollapalli, S. Kataria, H. Liu, and F. Xia, Eds., vol. WS-15-13. AAAI Press, 2015. [Online]. Available: <http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10185>
- [152] E. Tarnow, “First direct evidence of two stages in free recall,” *RUDN Journal of Psychology and Pedagogics*, no. 4, pp. 15–26, 2015.
- [153] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” 2020.
- [154] Y. Bernstein and J. Zobel, “Redundant documents and search effectiveness,” in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 736–743.
- [155] P. Bysani, “Detecting novelty in the context of progressive summarization,” in *Proceedings of the NAACL HLT 2010 Student Research Workshop*. Association for Computational Linguistics, 2010, pp. 13–18.
- [156] B. Gipp, N. Meuschke, and C. Breiteringer, “Citation-based plagiarism detection: Practicability on a large-scale scientific corpus,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 8, pp. 1527–1540, 2014.
- [157] V. Bhatnagar, A. S. Al-Hegami, and N. Kumar, “Novelty as a measure of interestingness in knowledge discovery,” *Constraints*, vol. 9, p. 18, 2006.
- [158] Y. Qin, D. Wurzer, V. Lavrenko, and C. Tang, “Spotting rumors via novelty detection,” *CoRR*, vol. abs/1611.06322, 2016. [Online]. Available: <http://arxiv.org/abs/1611.06322>

- 
- [159] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," in *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, 2003, pp. 314–321. [Online]. Available: <http://doi.acm.org/10.1145/860435.860493>
  - [160] Y. Zhang, J. P. Callan, and T. P. Minka, "Novelty and redundancy detection in adaptive filtering," in *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, 2002, pp. 81–88. [Online]. Available: <http://doi.acm.org/10.1145/564376.564393>
  - [161] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.
  - [162] R. Mihalcea and P. Tarau, "Texttrank: Bringing order into texts." Association for Computational Linguistics, 2004.
  - [163] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
  - [164] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.
  - [165] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1746–1751. [Online]. Available: <http://aclweb.org/anthology/D/D14/D14-1181.pdf>
  - [166] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2015, pp. 632–642. [Online]. Available: <http://aclweb.org/anthology/D/D15/D15-1075.pdf>
  - [167] L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, and Z. Jin, "Natural language inference by tree-based convolution and heuristic matching," *arXiv preprint arXiv:1512.08422*, 2015.
  - [168] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *HLT-NAACL*, vol. 13, 2013, pp. 746–751.
  - [169] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.
  - [170] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 2249–2255. [Online]. Available: <http://aclweb.org/anthology/D/D16/D16-1244.pdf>
  - [171] A. Lai, Y. Bisk, and J. Hockenmaier, "Natural language inference from multiple premises," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, 2017, pp. 100–109. [Online]. Available: <https://aclanthology.info/papers/I17-1011/i17-1011>
  - [172] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: <http://arxiv.org/abs/1409.0473>
  - [173] Y. Liu, C. Sun, L. Lin, and X. Wang, "Learning natural language inference using bidirectional LSTM model and inner-attention," *CoRR*, vol. abs/1605.09090, 2016. [Online]. Available: <http://arxiv.org/abs/1605.09090>
  - [174] D. Cer, Yinfei, Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.
  - [175] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.
  - [176] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," 2016.
  - [177] L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, and Z. Jin, "Recognizing entailment and contradiction by tree-based convolution," *CoRR*, vol. abs/1512.08422, 2015. [Online]. Available: <http://arxiv.org/abs/1512.08422>
  - [178] T. Dasgupta and L. Dey, "Automatic scoring for innovativeness of textual ideas," in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

- 
- [179] E. Tulving and N. Kroll, “Novelty assessment in the brain and long-term memory encoding,” *Psychonomic Bulletin & Review*, vol. 2, no. 3, pp. 387–390, 1995.
  - [180] M. J. Watkins and J. M. Gardiner, “An appreciation of generate-recognize theory of recall,” *Journal of Memory and Language*, vol. 18, no. 6, p. 687, 1979.
  - [181] G. Kumaran and J. Allan, “Text classification and named entities for new event detection,” in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004, pp. 297–304.
  - [182] K. W. Ng, F. S. Tsai, L. Chen, and K. C. Goh, “Novelty detection for text documents using named entity recognition,” in *Information, Communications & Signal Processing, 2007 6th International Conference on*. IEEE, 2007, pp. 1–5.
  - [183] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, “From word embeddings to document distances,” in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 957–966. [Online]. Available: <http://jmlr.org/proceedings/papers/v37/kusnerb15.html>
  - [184] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.
  - [185] F. S. Tsai and Y. Zhang, “D2S: document-to-sentence framework for novelty detection,” *Knowl. Inf. Syst.*, vol. 29, no. 2, pp. 419–433, 2011. [Online]. Available: <https://doi.org/10.1007/s10115-010-0372-2>
  - [186] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, “Hotpotqa: A dataset for diverse, explainable multi-hop question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018, pp. 2369–2380. [Online]. Available: <https://doi.org/10.18653/v1/d18-1259>
  - [187] I. Dagan, O. Glickman, and B. Magnini, “The PASCAL recognising textual entailment challenge,” in *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, ser. Lecture Notes in Computer Science, J. Q. Candela, I. Dagan, B. Magnini, and F. d’Alché-Buc, Eds., vol. 3944. Springer, 2005, pp. 177–190. [Online]. Available: [https://doi.org/10.1007/11736790\\_9](https://doi.org/10.1007/11736790_9)
  - [188] M. B.-D. I. D. H. G. D. Bentivogli, L., “The Sixth PASCAL Recognizing Textual Entailment Challenge,” in *Proceedings of the Text Analysis Conference (TAC 2010), November 15-16, 2010 National Institute of Standards and Technology Gaithersburg, Maryland, USA.*, 2010.
  - [189] C. P.-D. I. D. H. T. G. D. Bentivogli, L., “The Seventh PASCAL Recognizing Textual Entailment Challenge,” in *In TAC 2011 Notebook Proceedings, November 14-15, 2011, Gaithersburg, Maryland, USA.*, 2011.
  - [190] H. Trivedi, H. Kwon, T. Khot, A. Sabharwal, and N. Balasubramanian, “Repurposing entailment for multi-hop question answering tasks,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 2948–2958. [Online]. Available: <https://doi.org/10.18653/v1/n19-1302>
  - [191] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, “Enhanced LSTM for Natural Language Inference,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1657–1668. [Online]. Available: <https://www.aclweb.org/anthology/P17-1152>
  - [192] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. [Online]. Available: <https://www.aclweb.org/anthology/D15-1075>
  - [193] A. Williams, N. Nangia, and S. Bowman, “A broad-coverage challenge corpus for sentence understanding through inference,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 1112–1122. [Online]. Available: <http://aclweb.org/anthology/N18-1101>
  - [194] S. Burrows, M. Potthast, and B. Stein, “Paraphrase acquisition via crowdsourcing and machine learning,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 3, p. 43, 2013.



- 
- [195] J. F. Sánchez-Vega, “Identificación de plagio parafraseado incorporando estructura, sentido y estilo de los textos,” Ph.D. dissertation, PhD thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, 2016.
  - [196] A. Barrón-Cedeño, M. Vila, M. A. Martí, and P. Rosso, “Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection,” *Comput. Linguistics*, vol. 39, no. 4, pp. 917–947, 2013. [Online]. Available: <https://doi.org/10.1162/COLLa.00153>
  - [197] P. D. Clough and M. Stevenson, “Developing a corpus of plagiarised short answers,” *Lang. Resour. Evaluation*, vol. 45, no. 1, pp. 5–24, 2011. [Online]. Available: <https://doi.org/10.1007/s10579-009-9112-1>
  - [198] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, “Universal sentence encoder for english,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, E. Blanco and W. Lu, Eds. Association for Computational Linguistics, 2018, pp. 169–174. [Online]. Available: <https://doi.org/10.18653/v1/d18-2029>
  - [199] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.
  - [200] T. Ghosal, V. Edithal, A. Ekbal, P. Bhattacharyya, S. S. S. K. Chivukula, and G. Tsatsaronis, “Is your document novel? let attention guide you. an attention based model for document level novelty detection,” *Natural Language Engineering*, vol. 1, no. 1, pp. 1–38, 2020.
  - [201] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. E. Peters, M. Schmitz, and L. Zettlemoyer, “Allennlp: A deep semantic natural language processing platform,” *CoRR*, vol. abs/1803.07640, 2018. [Online]. Available: <http://arxiv.org/abs/1803.07640>
  - [202] O. François, “Arbitrariness of peer review: A bayesian analysis of the nips experiment,” *arXiv preprint arXiv:1507.06411*, 2015.
  - [203] M. De Rond and A. N. Miller, “Publish or perish: bane or boon of academic life?” *Journal of Management Inquiry*, vol. 14, no. 4, pp. 321–329, 2005.
  - [204] R. Smith, “Strategies for coping with information overload,” 2010.
  - [205] M. Kovanis, R. Porcher, P. Ravaut, and L. Trinquart, “The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise,” *PLoS One*, vol. 11, no. 11, p. e0166387, 2016.
  - [206] T. Ghosal, R. Verma, A. Ekbal, S. Saha, and P. Bhattacharyya, “Investigating impact features in editorial pre-screening of research papers,” in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*, 2018, pp. 333–334. [Online]. Available: <https://doi.org/10.1145/3197026.3203910>
  - [207] H. Stolowy, “Letter from the editor: Why are papers desk rejected at european accounting review?” *European Accounting Review*, vol. 26, no. 3, pp. 411–418, 2017. [Online]. Available: <https://doi.org/10.1080/09638180.2017.1347360>
  - [208] T. Ghosal, R. Sonam, S. Saha, A. Ekbal, and P. Bhattacharyya, “Investigating domain features for scope detection and classification of scientific articles,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA), may 2018.
  - [209] S. S. Leopold, “Increased manuscript submissions prompt journals to make hard choices,” 2015.
  - [210] S. Rose, D. Engel, N. Cramer, and W. Cowley, “Automatic keyword extraction from individual documents,” *Text Mining*, pp. 1–20, 2010.
  - [211] L. Kaufman and P. J. Rousseeuw, “Partitioning around medoids (program pam),” *Finding groups in data: an introduction to cluster analysis*, pp. 68–125, 1990.
  - [212] H. Müller, A. Foncubierta-Rodriguez, C. Lin, and I. Eggel, “Determining the importance of figures in journal articles to find representative images,” in *SPIE Proceedings*, vol. 8674, 2013, p. 9.
  - [213] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
  - [214] C. Clark and S. Divvala, “Pdffigures 2.0: Mining figures from research papers,” in *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*. IEEE, 2016, pp. 143–152.
  - [215] C. A. Clark and S. K. Divvala, “Pdffigures 2.0: Mining figures from research papers,” in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016, Newark, NJ, USA, June 19 - 23, 2016*, 2016, pp. 143–152. [Online]. Available: <https://doi.org/10.1145/2910896.2910904>

- [216] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.
- [217] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4945–4949.
- [218] Y. Wang, M. Huang, L. Zhao *et al.*, “Attention-based lstm for aspect-level sentiment classification,” in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.
- [219] S. Saha, S. Mitra, and S. Kramer, “Exploring multiobjective optimization for multiview clustering,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 4, p. 44, 2018.
- [220] L. Bornmann and H. Daniel, “Reliability of reviewers’ ratings when using public peer review: a case study,” *Learned Publishing*, vol. 23, no. 2, pp. 124–131, 2010. [Online]. Available: <https://doi.org/10.1087/20100207>
- [221] J. Langford and M. Guzdial, “The arbitrariness of reviews, and advice for school administrators,” *Commun. ACM*, vol. 58, no. 4, pp. 12–13, 2015. [Online]. Available: <https://doi.org/10.1145/2732417>
- [222] R. Smith, “Peer review: a flawed process at the heart of science and journals,” *Journal of the royal society of medicine*, vol. 99, no. 4, pp. 178–182, 2006.
- [223] A. Tomkins, M. Zhang, and W. D. Heavlin, “Reviewer bias in single- versus double-blind peer review,” *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 48, pp. 12 708–12 713, 2017. [Online]. Available: <https://doi.org/10.1073/pnas.1707323114>
- [224] N. B. Shah, B. Tabibian, K. Muandet, I. Guyon, and U. Von Luxburg, “Design and analysis of the nips 2016 review process,” *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1913–1946, 2018.
- [225] E. S. Brezis and A. Birukou, “Arbitrariness in the peer review process,” *Scientometrics*, pp. 1–19, 2020.
- [226] H. Ledford and R. V. Noorden, “High-profile coronavirus retractions raise concerns about data oversight,” <https://www.nature.com/articles/d41586-020-01695-w>, June 2020, (Accessed on 09/02/2020).
- [227] K. Cohen, K. Fort, M. Mieskes, and A. Névéol, “Reviewing natural language processing research,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Online: Association for Computational Linguistics, Jul. 2020, pp. 16–18. [Online]. Available: <https://www.aclweb.org/anthology/2020.acl-tutorials.4>
- [228] EMNLP, “Advice on reviewing for emnlp,” <https://2020.emnlp.org/blog/2020-05-17-write-good-reviews>, A 2020, (Accessed on 09/02/2020).
- [229] J. Huisman and J. Smits, “Duration and quality of the peer review process: the author’s perspective,” *Scientometrics*, vol. 113, no. 1, pp. 633–650, 2017.
- [230] L. Bornmann and H.-D. Daniel, “Reliability of reviewers’ ratings when using public peer review: a case study,” *Learned Publishing*, vol. 23, no. 2, pp. 124–131, 2010.
- [231] J. Langford and M. Guzdial, “The arbitrariness of reviews, and advice for school administrators,” *Communications of the ACM*, vol. 58, no. 4, pp. 12–13, 2015.
- [232] A. Rogers, “Peer review in nlp: reject-if-not-sota,” <https://hackingsemantics.xyz/2020/reviewing-models/>, April 2020, (Accessed on 09/02/2020).
- [233] K. Shashok, “Content and communication: How can peer review provide helpful feedback about the writing?” *BMC Medical Research Methodology*, vol. 8, no. 1, p. 3, 2008.
- [234] A. Olena, “How to make scientists into better peer reviewers,” <https://www.the-scientist.com/careers/how-to-make-scientists-into-better-peer-reviewers-30>, 2018.
- [235] D. Sculley, J. Snoek, and A. Wiltschko, “Avoiding a tragedy of the commons in the peer review process,” *arXiv preprint arXiv:1901.06246*, 2018.
- [236] S. Price and P. A. Flach, “Computational support for academic peer review: a perspective from artificial intelligence.” *Commun. ACM*, vol. 60, no. 3, pp. 70–79, 2017.
- [237] D. Heaven, “Ai peer reviewers unleashed to ease publishing grind,” <https://www.nature.com/articles/d41586-018-07245-9>, 2018.
- [238] C. H. Davis, B. L. Bass, K. E. Behrns, K. D. Lillemoe, O. J. Garden, M. S. Roh, J. E. Lee, C. M. Balch, and T. A. Aloia, “Reviewing the review: a qualitative assessment of the peer review process in surgical journals.” *Research integrity and peer review*, vol. 3, no. 1, p. 4, 2018.

- 
- [239] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
  - [240] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” *arXiv preprint arXiv:1903.10676*, 2019.
  - [241] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, “Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks,” *arXiv preprint arXiv:2010.08240*, 10 2020. [Online]. Available: <https://arxiv.org/abs/2010.08240>
  - [242] C. Xiong, V. Zhong, and R. Socher, “Dynamic coattention networks for question answering,” *arXiv preprint arXiv:1611.01604*, 2016.
  - [243] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.
  - [244] D. Pride and P. Knuth, “Incidental or influential? - A decade of using text-mining for citation function classification,” in *Proceedings of the 16th International Conference on Scientometrics and Informetrics, ISSI 2017, Wuhan, China, October 16-20, 2017*, J. Qiu, R. Rousseau, C. R. Sugimoto, and F. Xin, Eds. ISSI Society, 2017, pp. 1357–1367.
  - [245] R. Rousseau, “The influence of missing publications on the hirsch index,” *Journal of Informetrics*, vol. 1, no. 1, pp. 2–7, 2007.
  - [246] R. Van Noorden, “The science that’s never been cited,” *Nature*, vol. 552, 2017.
  - [247] R. West, K. Stenius, and T. Kettunen, “Use and abuse of citations,” *Addiction Science: A Guide for the Perplexed*, p. 191, 2017.
  - [248] G.-A. Viiu, “A theoretical evaluation of hirsch-type bibliometric indicators confronted with extreme self-citation,” *Journal of Informetrics*, vol. 10, no. 2, pp. 552–566, 2016.
  - [249] R. Van Noorden and D. Singh Chawla, “Hundreds of extreme self-citing scientists revealed in new database,” *Natur*, vol. 572, no. 7771, pp. 578–579, 2019.
  - [250] A. W. Wilhite and E. A. Fong, “Coercive citation in academic publishing,” *Science*, vol. 335, no. 6068, pp. 542–543, 2012.
  - [251] X. Bai, F. Xia, I. Lee, J. Zhang, and Z. Ning, “Identifying anomalous citations for objective evaluation of scholarly article impact,” *PloS one*, vol. 11, no. 9, p. e0162364, 2016.
  - [252] C. Bartneck and S. Kokkermans, “Detecting h-index manipulation through self-citation analysis,” *Scientometrics*, vol. 87, no. 1, pp. 85–98, 2011.
  - [253] G. A. Ronda-Pupo and T. Pham, “The evolutions of the rich get richer and the fit get richer phenomena in scholarly networks: the case of the strategic management journal,” *Scientometrics*, vol. 116, no. 1, pp. 363–383, 2018.
  - [254] M. Camacho-Miñano and M. Núñez-Nickel, “The multilayered nature of reference selection,” *J. Assoc. Inf. Sci. Technol.*, vol. 60, no. 4, pp. 754–777, 2009. [Online]. Available: <https://doi.org/10.1002/asi.21018>
  - [255] J. H. C. Cerdá, E. M. Nieto, and M. L. Campos, “What’s wrong with citation counts?” *D-Lib Magazine*, vol. 15, no. 3/4, pp. 1082–9873, 2009.
  - [256] F. Laloë and R. Mosseri, “Bibliometric evaluation of individual researchers: not even right... not even wrong!” *Europhysics News*, vol. 40, no. 5, pp. 26–29, 2009.
  - [257] P. Lopez, “Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications,” in *International conference on theory and practice of digital libraries*. Springer, 2009, pp. 473–474.
  - [258] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, “Yake! collection-independent automatic keyword extractor,” in *European Conference on Information Retrieval*. Springer, 2018, pp. 806–810.
  - [259] G. Huang, C. Guo, M. J. Kusner, Y. Sun, F. Sha, and K. Q. Weinberger, “Supervised word mover’s distance,” in *Advances in Neural Information Processing Systems*, 2016, pp. 4862–4870.

- 
- [260] C. Gilbert and E. Hutto, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Available at (20/04/16) <http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>, vol. 81, 2014, p. 82.
- [261] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [262] S. Nazir, M. Asif, and S. Ahmad, "Important citation identification by exploiting the optimal in-text citation frequency," in *2020 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2020, pp. 1–6.
- [263] S. Nazir, M. Asif, S. Ahmad, F. Bukhari, M. T. Afzal, and H. Aljuaid, "Important citation identification by exploiting content and section-wise in-text citation count," *PloS one*, vol. 15, no. 3, p. e0228885, 2020.
- [264] D. Pride and P. Knoch, "Incidental or influential? - challenges in automatically detecting citation importance using publication full texts," in *Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings*, ser. Lecture Notes in Computer Science, J. Kamps, G. Tsakonas, Y. Manolopoulos, L. S. Iliadis, and I. Karydis, Eds., vol. 10450. Springer, 2017, pp. 572–578. [Online]. Available: [https://doi.org/10.1007/978-3-319-67008-9\\_48](https://doi.org/10.1007/978-3-319-67008-9_48)

# Publications

---

## Accepted Journal Publication

1. **Ghosal, T.**, Edithal, V., Ekbal, A., Bhattacharyya, P., Tsatsaronis, G., & Chivukula, S. S. S. K. (2019), *Is your document novel? Let attention guide you. An Attention-based model for Document-level Novelty Detection*, Natural Language Engineering, pp. 1–28, 2020.

## Accepted Conference Publications

1. **Ghosal, T.**, Salam, A., Tiwari, S., Ekbal, A., & Bhattacharyya, P. (2018). *TAP-DLND 1.0: A Corpus for Document Level Novelty Detection*. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.
2. **Ghosal, T.**, Edithal, V., Ekbal, A., Bhattacharyya, P., Tsatsaronis, G., & Chivukula, S. S. S. K. (2018, August). *Novelty Goes Deep. A Deep Neural Solution To Document Level Novelty Detection*. In Proceedings of the 27th International Conference on Computational Linguistics (COLING) (pp. 2802-2813).
3. **Ghosal, T.**, Shukla, A., Ekbal, A., & Bhattacharyya, P. (2019, July). *To Comprehend the New: On Measuring the Freshness of a Document*. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
4. **Ghosal, T.**, Sonam, R., Ekbal, A., Saha, S., & Bhattacharyya, P. (2019, June). *Is the Paper Within Scope? Are You Fishing in the Right Pond?*. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 237-240). IEEE.
5. **Ghosal, T.**, Raj, A., Ekbal, A., Saha, S., & Bhattacharyya, P. (2019, June). *A Deep Multimodal Investigation To Determine the Appropriateness of Scholarly Submissions*. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 227-236). IEEE.
6. **Ghosal, T.**, Dey, D., Dutta, A., Ekbal, A., Saha, S., & Bhattacharyya, P. (2019, June). *A Multiview Clustering Approach To Identify Out-of-Scope Submissions in Peer Review*. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 392-393). IEEE.
7. **Ghosal, T.**, Verma, R., Ekbal, A., & Bhattacharyya, P. (2019, July). *DeepSentiPeer: Harnessing Sentiment in Review Texts to Recommend Peer Review Decisions*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 1120-1130).

8. **Ghosal, T.**, Verma, R., Ekbal, A., Saha, S., & Bhattacharyya, P. (2020, November). *An Empirical Study of Importance of Different Sections in Research Articles Towards Ascertaining Their Appropriateness to a Journal*. In Proceedings of the 22nd International Conference on Asian Digital Libraries (ICADL) (pp. 407-415).

### Under Communication Publications

1. **Ghosal, T.**, Verma, R., Ekbal, A., & Bhattacharyya, P. (2020). *Not All Reviews Are Equal. Can You Identify Significant Peer Reviews?* [Under Communication]
2. **Ghosal, T.**, Biswas, T., Saikh, T., Ekbal, A., & Bhattacharyya, P. (2020). *Textual Novelty Detection: An NLP Perspective* [Under Communication]
3. **Ghosal, T.**, Patton, R., Stahl, C., Ekbal, A., & Bhattacharyya, P. (2020). *Establishing a Research Lineage via Identification of Meaningful Citations* [Under Communication]

### Other Related Publications

1. **Ghosal, T.**, Verma, R., Ekbal, A., & Bhattacharyya, P. (2019, June). *A Sentiment Augmented Deep Architecture to Predict Peer Review Outcomes*. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 414-415). IEEE.
2. **Ghosal, T.**, Sonam, R., Saha, S., Ekbal, A., & Bhattacharyya, P. (2018). *Investigating domain features for scope detection and classification of scientific articles*. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (pp. 7-12).
3. **Ghosal, T.**, Chakraborty, A., Sonam, R., Ekbal, A., Saha, S., & Bhattacharyya, P. (2019, June). *Incorporating Full Text and Bibliographic Features to Improve Scholarly Journal Recommendation*. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 374-375). IEEE.
4. **Ghosal, T.**, Verma, R., Ekbal, A., Saha, S., & Bhattacharyya, P. (2018, May). *Investigating Impact Features in Editorial Pre-Screening of Research Papers*. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (pp. 333-334). ACM.
5. **Ghosal, T.** (2018) *Exploring the Implications of Artificial Intelligence in Various Aspects of Scholarly Peer Review*. Bulletin of the IEEE Technical Committee on Digital Libraries , In the Doctoral Consortium of the 18th ACM/IEEE Joint Conference on Digital Libraries
6. Saikh, T., **Ghosal, T.**, Ekbal, A., & Bhattacharyya, P. (2017, December). *Document Level Novelty Detection: Textual Entailment Lends a Helping Hand*. In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017) (pp. 131-140).