NLP and IR for Scholarly Data Processing

Tirthankar Ghosal
Indian Institute of Technology Patna
tirthankar.slg@gmail.com
https://tirthankarslg.wixsite.com/ainlpmldl
@TirthankarSlg

Introduction

- A fifth-year PhD scholar from Indian Institute of Technology (IIT) Patna

- AI-NLP-ML Lab, IIT Patna

  http://www.iitp.ac.in/~ai-nlp-ml/

- Research area: NLP, Machine Learning, Deep Learning, Information Extraction and Retrieval on Scholarly Data, Digital Libraries, Bibliometric Intelligence

- Supervisors:

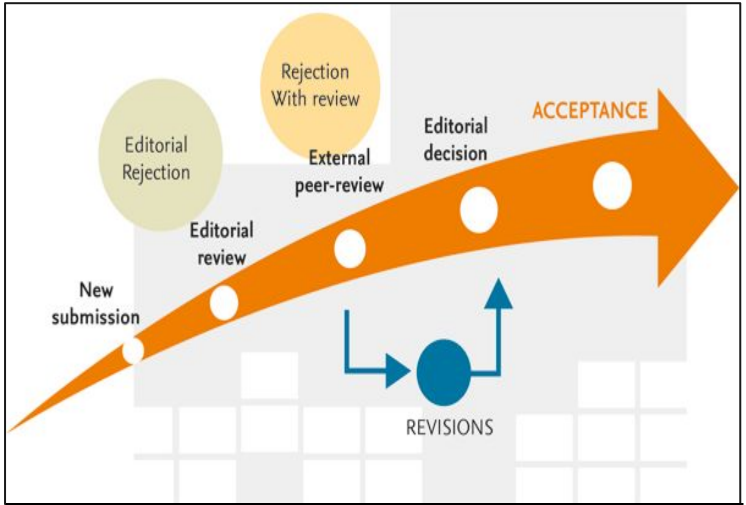
  Dr. Asif Ekbal (IIT Patna) & Prof. Pushpak Bhattacharyya (IIT Bombay and IIT Patna)

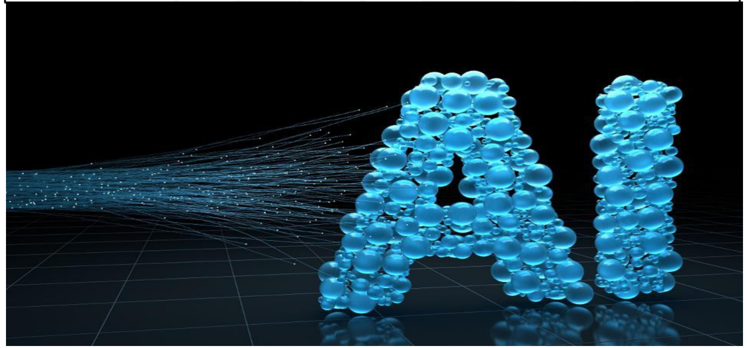
Brief Curriculum Vitae

- Pursuing Ph.D. at Indian Institute of Technology Patna, India since January 2016
- Visvesvaraya Research Fellow
- Assistant Professor of Computer Science and Engineering at Sikkim Manipal Institute of Technology, Sikkim Manipal University, India (2012-2015)
- **https://tirthankarslg.wixsite.com/ainlpmldl**

Ph.D. Summary

- Journey started in January 2016
- Exploring 3 broad problems associated with Scholarly Peer Review
  - Document-Level Novelty Detection
  - Finding Appropriateness of an Article to a Journal (Scope Detection)
  - Quality Assessment of Articles (post-publication): at ORNL
  - AI in the Peer Review pipeline
- *So many interesting problems to explore !! But so difficult the data is …*
  - To mine
  - To obtain
  - Too intelligent !!
- AWSAR Story: https://www.awsar-dst.in/assets/winner_article_2018/43_PhD.pdf





Why do we have the scholarly ecosystem?

NLP and IR for processing Scholarly Knowledge

1. Why do we want to mine scholarly knowledge?

- Manifestation of highest form of human intelligence
- Vast knowledge remain **under-processed**
- But texts are not simple: they are **intelligent**

2. Synergy: **Meta Science**

Natural Language Processing

Machine Learning

Digital Libraries

Information Retrieval

Knowledge Discovery

Scientometrics/Bibliometric Intelligence

AI for Peer Review: Mining Scholarly Data

Peer Review: the cornerstone of modern scientific progress. (Is it?)

What are the challenges of current peer review system?

- *We are doing research in the 21st century with validating techniques of the 16th century*
- *Exponential Rise of Research Articles. Is Science making progress at an exponential rate?*
- *Time-consuming*
- *Biased*
- *How fragile our peer review system is?*
- *Predatory journals*
- *Ghost peer reviewers*
- *Bad quality of reviews*
- *Finding relevant prior knowledge*
- *Coercive citations*

Motivation

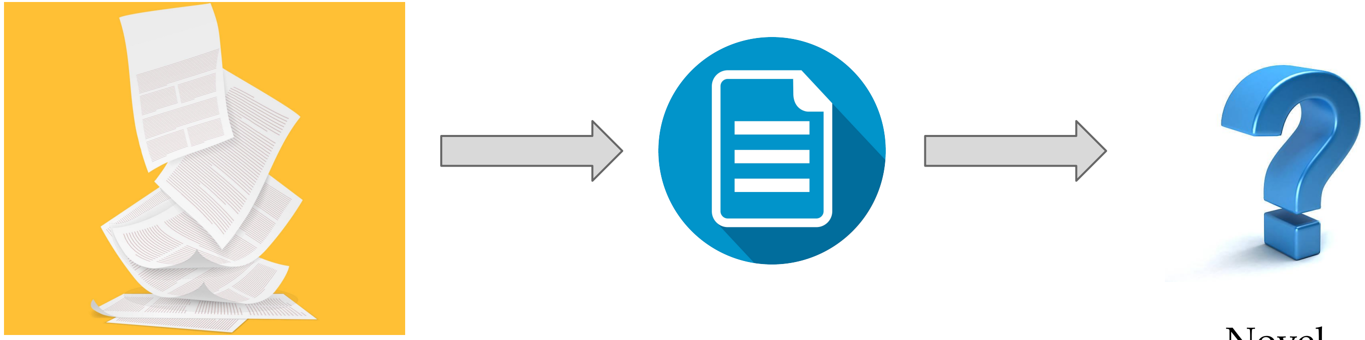
- Exponential rise of redundant/duplicate information/documents across the web [Big Data]
- Redundancy at the semantic level (text reuse,rewrite,paraphrase, etc.) [NLP]
  - Existing methods are lexical,IR-oriented,rule-based, operating at the sentence-level
- Plagiarism Detection at the semantic and pragmatic level [Curb predatory publishing]
- Assist the editors to efficiently identify out-of-Scope papers; speed up the peer review process - flag out irrelevant submissions [Desk Rejection]
- Assess the impact of an article post publication; research pervasiveness; research lineage; faster relevant literature discovery [Quality]
- An AI empowered to quest for new knowledge [Scientific Novelty]

Novelty Detection

- ***Novelty:*** *The search of new; eternal quest of the inquisitive mind*

*Of all the passions that possess mankind, / The love of **novelty** most rules the mind; / In search of this, from realm to realm we roam, / Our fleets come fraught with every folly home.*

-Shelby Foote



Novel

Textual Novelty Detection

- Novelty Mining: elicit new information from texts
- An IR task for long: retrieve novel sentences
- Document-Level Novelty Detection: A frontier less explored
- Properties (Ghosal et. al, 2018):
  - Relevance
  - Relativity
  - Diversity
  - Temporality
- Applications in diverse domains of information processing :
  - Extractive text summarization
  - News Tracking
  - Predicting impact of scholarly articles

{ w.r.t. a set of seed documents called as the source *or* information already known/memory of the reader }

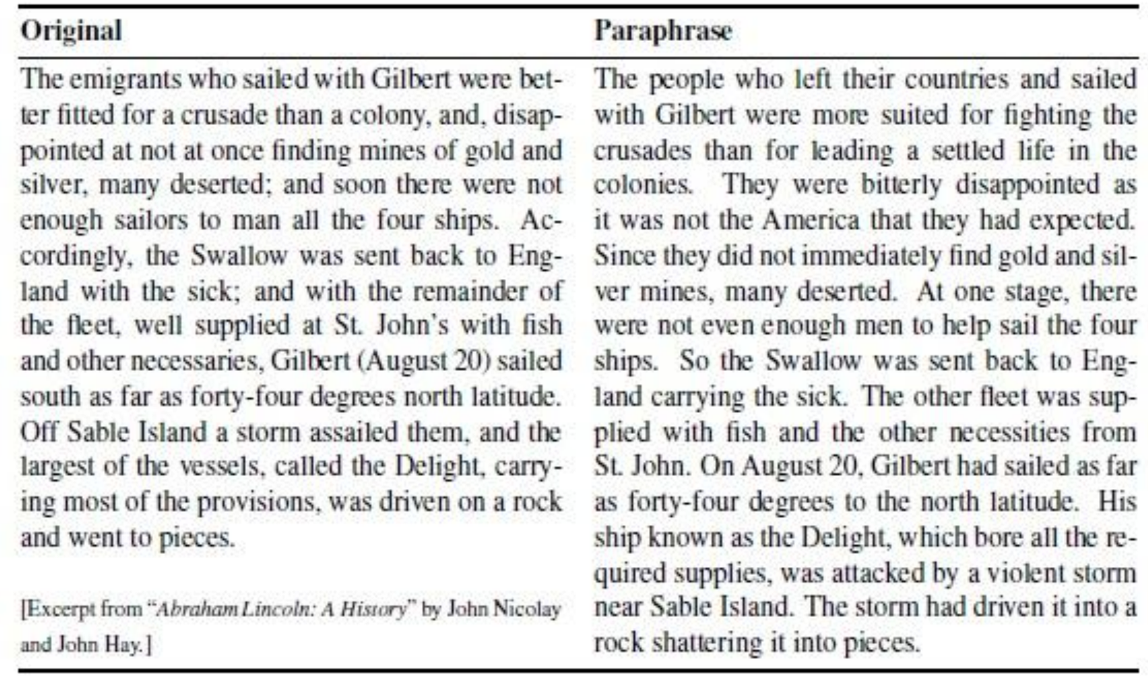
TAP-DLND 1.0: A Corpus for Document Level Novelty Detection by **Tirthankar Ghosal**, Amitra Salam, Swati Tiwari, Asif Ekbal and Pushpak Bhattacharyya published as a full paper at LREC 2018 held at Miyazaki, Japan

Problem Definition

- Categorize a document as novel or non-novel based on sufficient relevant new information
- For e.g., :
  - *d1 : Singapore is an island city-state located at the southern tip of the Malay Peninsula. It lies 137 kilometers north of the equator.*
  - *d2 : Singapore's territory consists of one main island along with 62 other islets. The population in Singapore is 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.*
- If we consider source as d1 and d2; d3 is novel, d4 is non-novel
- We take a very objective and simplistic view considering only the new information content.

TAP-DLND 1.0: A Corpus for Document Level Novelty Detection by **Tirthankar Ghosal**, Amitra Salam, Swati Tiwari, Asif Ekbal and Pushpak Bhattacharyya published as a full paper at LREC 2018 held at Miyazaki, Japan

Document-Level Redundancy/Non-Novelty



Original	Paraphrase
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 necessaries, 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. [Excerpt from "*Abraham Lincoln: A History*" by John Nicolay and John Hay.]	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.

❖ Non-Novel (https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/corpora/corpus-webis-cpc-11/)

★ We investigate whether a deep network can be trained to perceive novelty at the document-level and also identify semantically redundant/non-novel documents

Investigation Line

Can we train a neural network to predict novelty scores along this line to reach as close as possible to the ground truth? Classify a document?

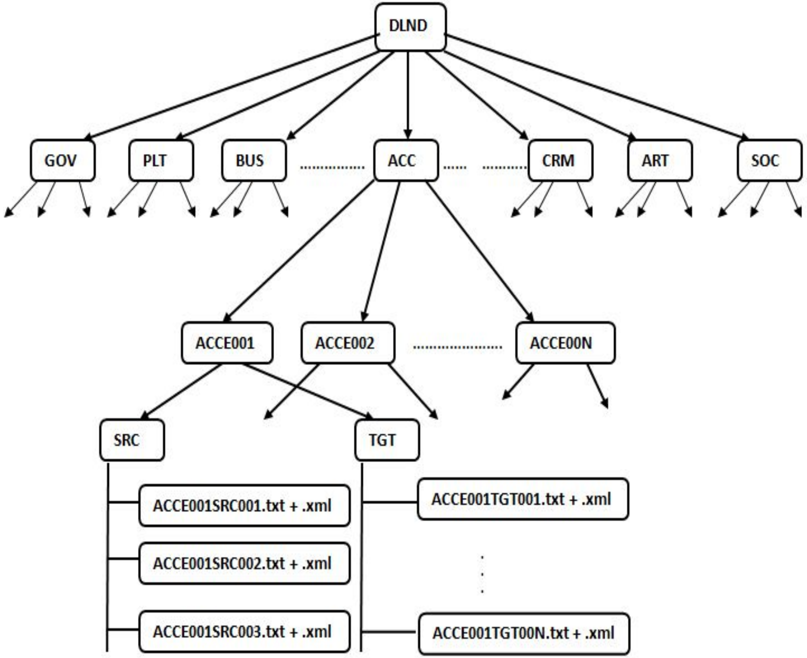
We decided to build a dataset manifesting the four general properties of novelty as we outline in (Ghosal et al. 2018) and the quantification as showed in the previous example.

➔ Relevance (source and target should be relevant; jaguar vs jaguar vs jaguar)
➔ Relativity (amount of NEWNESS?)
➔ Diversity (diverse information; not known previously)
➔ Temporality (novelty is usually a temporal update over existing knowledge)

TAP-DLND 1.0: A Corpus for Document Level Novelty Detection by **Tirthankar Ghosal**, Amitra Salam, Swati Tiwari, Asif Ekbal and Pushpak Bhattacharyya published as a full paper at LREC 2018 (H-Index: 56) held at Miyazaki, Japan

Dataset: TAP-DLND 1.0

- TAP-DLND 1.0 (Tirthankar-Asif-Pushpak Document-Level Novelty Detection Corpus)
  - A balanced document-level novelty detection dataset
  - Consists of events belonging to different categories
  - Satisfying Relevance, Relativity, Diversity, Temporality criteria for Novelty
  - 3 source documents per event; target documents are annotated against the information contained in the source documents
  - Binary Classification: Novel or Non-Novel
  - 2736 novel and 2704 non-novel documents
  - Inter-annotator agreement is 0.82



TAP-DLND 1.0 Structure

TAP-DLND 1.0: A Corpus for Document Level Novelty Detection by **Tirthankar Ghosal**, Amitra Salam, Swati Tiwari, Asif Ekbal and Pushpak Bhattacharyya published as a full paper at LREC 2018 (H-Index: 56) held at Miyazaki, Japan

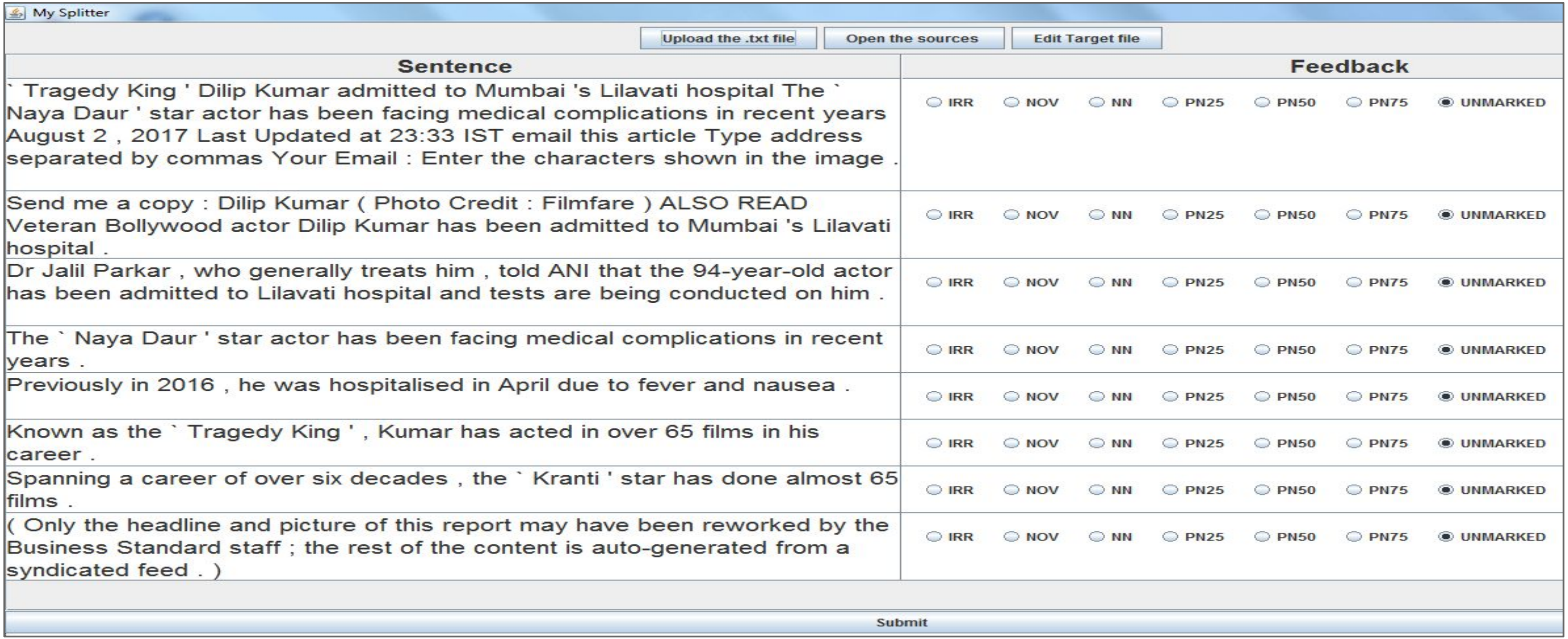
Dataset: TAP-DLND 1.1

- Document-level annotations were much too subjective.
- Annotate at the sentence-level.
- We extend the dataset. Include 2000 more target documents.
- Sentence-Level annotations gave us a document-level novelty score.
- Average of all sentence scores

Dataset Characteristics	**Statistics**
Event Categories	10
# Events	245
# Source documents/event	3
Total target documents	7536
Total sentences annotated	120,116
Avg sentences/document	~16
Avg words/document	~385
Inter-rater agreement (Kappa)	0.88

To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

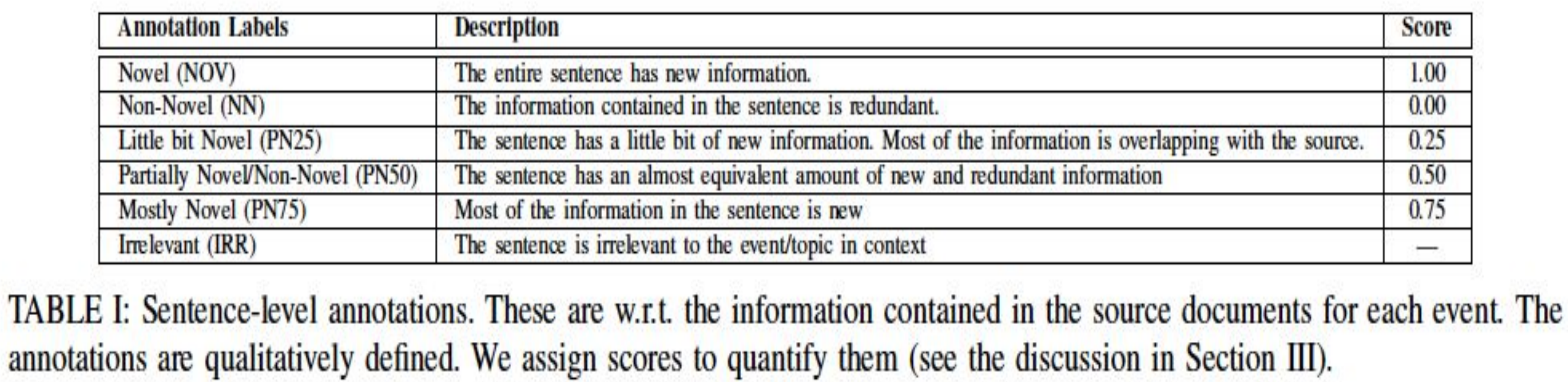
Annotation Interface



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 .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● 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 .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● 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 .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
The ` Naya Daur ' star actor has been facing medical complications in recent years .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
Previously in 2016 , he was hospitalised in April due to fever and nausea .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
Known as the ` Tragedy King ' , Kumar has acted in over 65 films in his career .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
Spanning a career of over six decades , the ` Kranti ' star has done almost 65 films .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● 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 .)	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED

To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

Annotation Labels



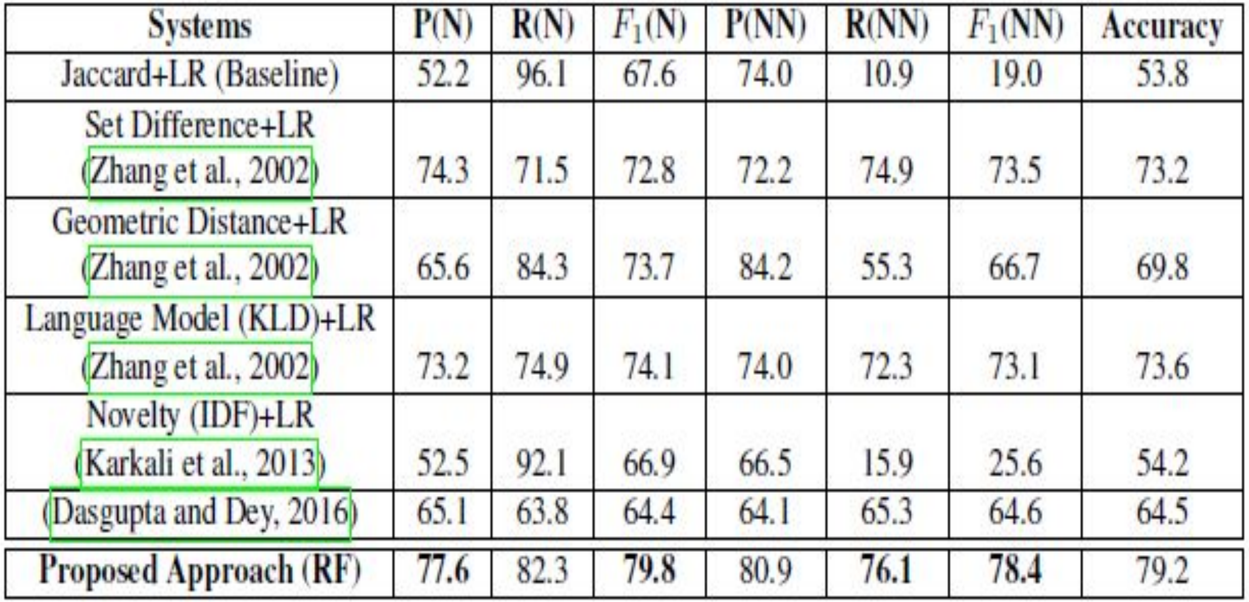
Annotation Labels	Description	Score
Novel (NOV)	The entire sentence has new information.	1.00
Non-Novel (NN)	The information contained in the sentence is redundant.	0.00
Little bit Novel (PN25)	The sentence has a little bit of new information. Most of the information is overlapping with the source.	0.25
Partially Novel/Non-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	—

TABLE I: Sentence-level annotations. These are w.r.t. the information contained in the source documents for each event. The annotations are qualitatively defined. We assign scores to quantify them (see the discussion in Section III).

To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

Feature-Based Solution (Model-I)

1. Semantic Similarity (Doc2Vec+Cosine)
2. Concept Centrality (TextRank, word2vec average+cosine
3. N-grams Similarity (n=2,3,8)
4. Named Entities and Keywords Match
5. New Word Count
6. Divergence (Language Model)



Systems	P(N)	R(N)	F_1(N)	P(NN)	R(NN)	F_1(NN)	Accuracy
Jaccard+LR (Baseline)	52.2	96.1	67.6	74.0	10.9	19.0	53.8
Set Difference+LR (Zhang et al., 2002)	74.3	71.5	72.8	72.2	74.9	73.5	73.2
Geometric Distance+LR (Zhang et al., 2002)	65.6	84.3	73.7	84.2	55.3	66.7	69.8
Language Model (KLD)+LR (Zhang et al., 2002)	73.2	74.9	74.1	74.0	72.3	73.1	73.6
Novelty (IDF)+LR (Karkali et al., 2013)	52.5	92.1	66.9	66.5	15.9	25.6	54.2
(Dasgupta and Dey, 2016)	65.1	63.8	64.4	64.1	65.3	64.6	64.5
Proposed Approach (RF)	**77.6**	82.3	**79.8**	80.9	**76.1**	**78.4**	79.2

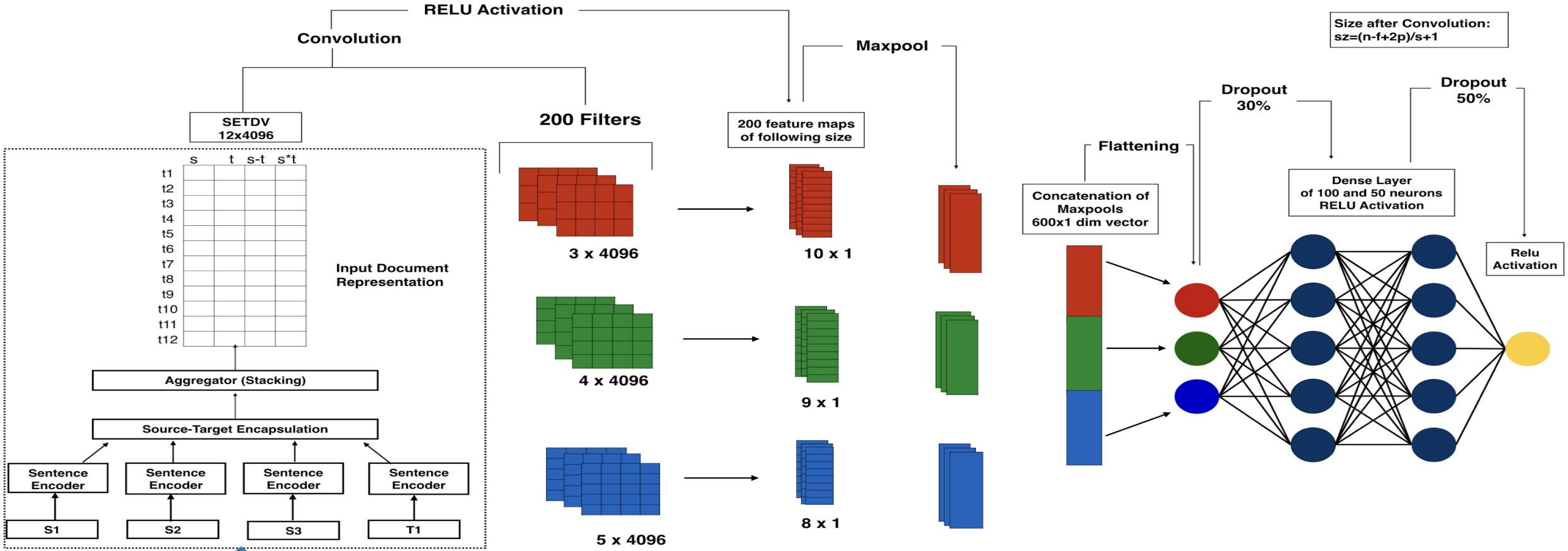
TAP-DLND 1.0: A Corpus for Document Level Novelty Detection by **Tirthankar Ghosal**, Amitra Salam, Swati Tiwari, Asif Ekbal and Pushpak Bhattacharyya published as a full paper at LREC 2018 (H-Index: 56) held at Miyazaki, Japan

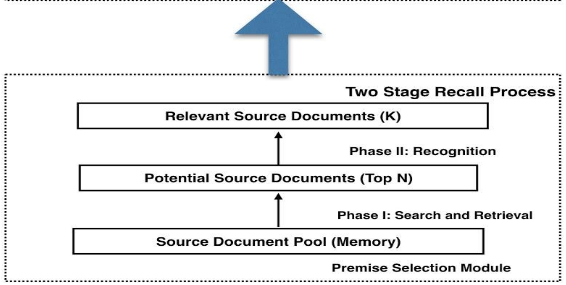
Deep Learning: Architecture Description (Model-II)

- Premise Selection (Approximating Two-Stage Theory of Human Recall to select the appropriate source documents for a given target document)
  - Phase-I: Search and Retrieval (Recall@10=0.93)
  - Phase-II: Recognition (Recall@3=0.94)
- Source-Encapsulated Target Document Vector (SETDV)
  - The nearest source sentence to one target sentence is selected via cosine similarity
  - The selected source and target sentence is encapsulated as:
    - t | s | t-s | t*s
  - The source encapsulated target sentence encodings are stacked to form the SETDV
  - The SETDV is fed to a Convolutional Neural Network (CNN) for feature extraction followed by a dense layer and finally a ReLU to predict the novelty score.

To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

Proposed Model-II




Two Stage Recall Process
Relevant Source Documents (K)
Phase II: Recognition
Potential Source Documents (Top N)
Phase I: Search and Retrieval
Source Document Pool (Memory)
Premise Selection Module

To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

Rationale

- *Our rationale behind the SETDV-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*.
- Relevance criteria is inherently manifested within the datasets we work with.
- The proposed architecture looks for relative, diverse new information of a target with respect to corresponding sources and learns the notion of a novel or non-novel document.
- Learning of novel vs. non-novel patterns via the relative representation

Novelty Goes Deep: A Deep Neural Approach Towards Document Level Novelty Detection by Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, Sameer Chivukula and Georgios Tsatsaronis accepted as a full paper in the 27th International Conference on Computational Linguistics (COLING 2018) held at Santa Fe, New-Mexico, USA (CORE rank A/H-Index: 43).

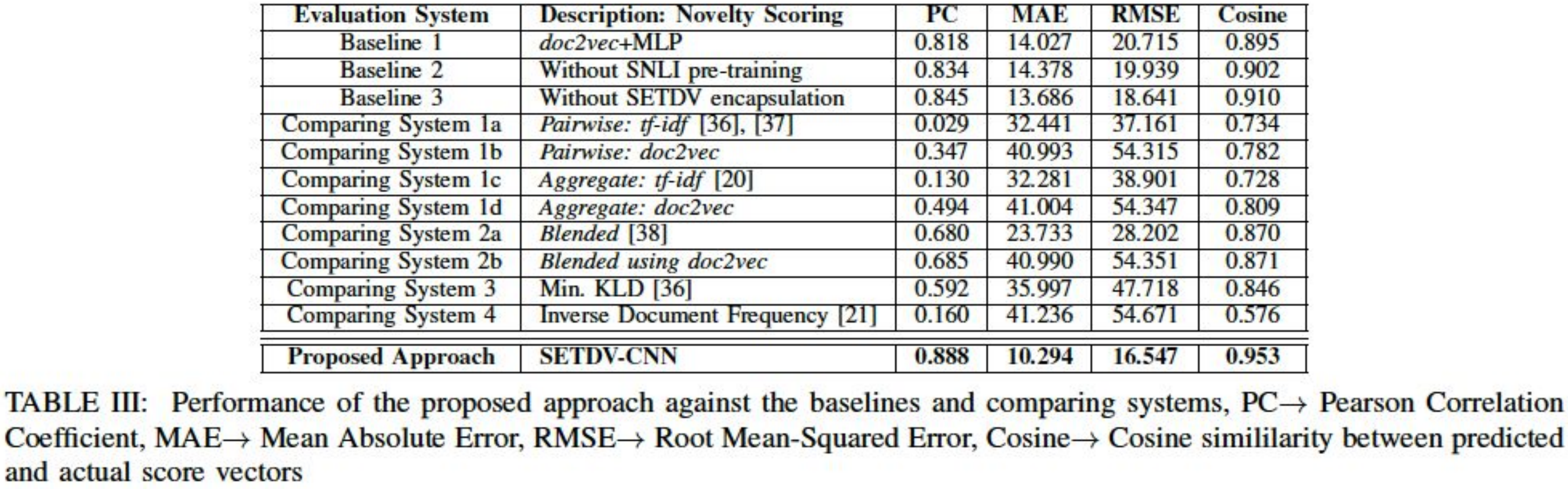
To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

Proposed Model-II (SETDV-CNN)

- Intuition: A non-novel document would contain many redundant sentences. Hence cosine similarity will pull that particular source sentence which contributes more towards making the target sentence redundant. Hence a joint encoding of source+redundant sentence would be different from that of a source+novel sentence.
- Thus the SETDV of a non-novel document would be different from that of a novel document.
- A Convolutional Neural Network (CNN) is then trained with the SETDV of the target documents.
- Finally the CNN extracted features are fed to a affine layer followed by a ReLU layer for final score prediction.

To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

Results (Novelty-Score Prediction)



Evaluation System	Description: Novelty Scoring	PC	MAE	RMSE	Cosine
Baseline 1	*doc2vec*+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	*Pairwise: tf-idf* [36], [37]	0.029	32.441	37.161	0.734
Comparing System 1b	*Pairwise: doc2vec*	0.347	40.993	54.315	0.782
Comparing System 1c	*Aggregate: tf-idf* [20]	0.130	32.281	38.901	0.728
Comparing System 1d	*Aggregate: doc2vec*	0.494	41.004	54.347	0.809
Comparing System 2a	*Blended* [38]	0.680	23.733	28.202	0.870
Comparing System 2b	*Blended using doc2vec*	0.685	40.990	54.351	0.871
Comparing System 3	Min. KLD [36]	0.592	35.997	47.718	0.846
Comparing System 4	Inverse Document Frequency [21]	0.160	41.236	54.671	0.576
Proposed Approach	**SETDV-CNN**	**0.888**	**10.294**	**16.547**	**0.953**

TABLE III: 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 simililarity between predicted and actual score vectors

Our baselines also served as a means of our ablation study. Baseline 1→ Without SNLI training and SETDV, Baseline 2→ Without SNLI pre-training of the sentence encoder, Baseline 3→ Without SETDV encapsulation

To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

Analysis

- ❖ We could see from the scatter plot that our approach closely approximates the ground-truth.
- ❖ It is clear from the results, that our proposed DNN outperforms the existing methods and baselines.
- ❖ Also ablation study shows that, the importance of each of the components towards the predictive capability of the DNN architecture

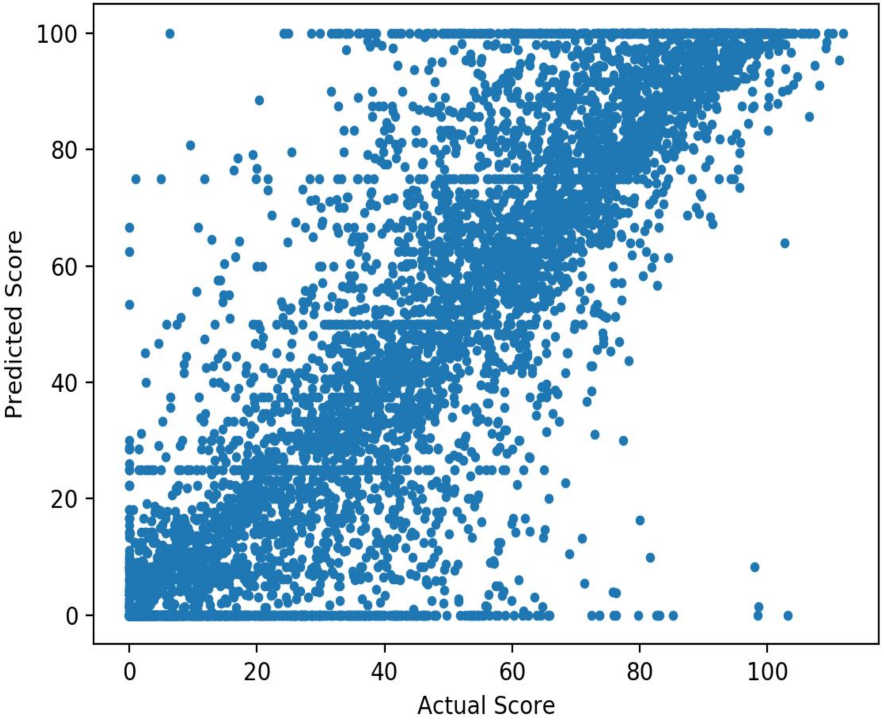
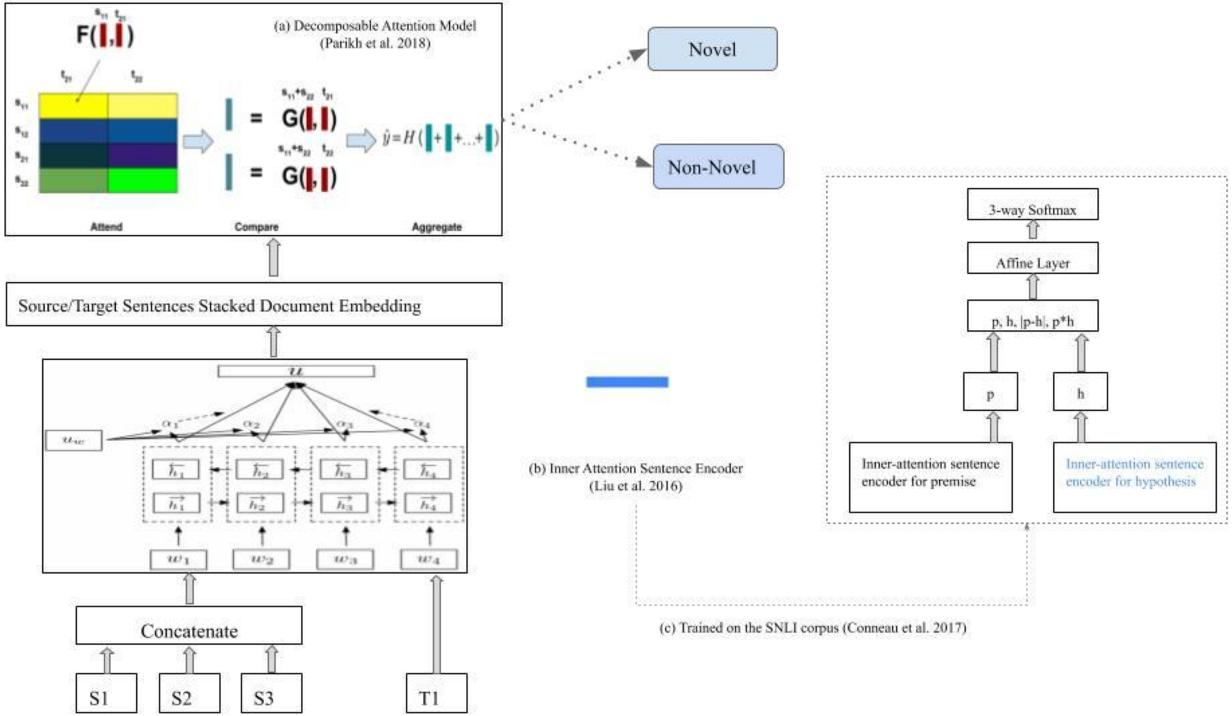


Figure: Scatter plot for predicted vs actual score

To Comprehend the New: On Measuring the Freshness of a Document by Tirthankar Ghosal, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks (IJCNN 2019) to be held at Budapest, Hungary (CORE rank A/H-Index: 41).

Deep Architecture (Model-III)

- Attention-based
- Order of less parameters than earlier models
- Efficient premise selection
- Leveraging natural language inference phenomena for novelty detection



Is Your Document Novel? Let Attention Guide You. An Attention-Based Model For Document Level Novelty Detection by Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal and Pushpak Bhattacharyya provisionally accepted in Natural Language Engineering (NLE) Journal by Cambridge University Press, 2019

Results on APWSJ dataset

- On the APWSJ dataset. Except the proposed methods we take all other numbers from (Zhang et al., 2002)
- Mistake=100-Accuracy as is there in the original paper.

Novelty Goes Deep: A Deep Neural Approach Towards Document Level Novelty Detection by Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, Sameer Chivukula and Georgios Tsatsaronis accepted as a full paper in the 27th International Conference on Computational Linguistics (COLING 2018) held at Santa Fe, New-Mexico, USA (CORE rank A/H-Index: 35).

Is Your Document Novel? Let Attention Guide You. An Attention-Based Model For Document Level Novelty Detection by Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal and Pushpak Bhattacharyya provisionally accepted in Natural Language Engineering (NLE) Journal by Cambridge University Press

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: Mixed	0.56	0.67	27.4%
Proposed Method (RDV-CNN)	**0.58**	**0.76**	**22.9%**
Proposed Method (Dec_Attn)	**0.86**	**0.92**	**7.8%**

Results on the Paraphrase Detection Task

- On the Webis-CPC-11 dataset
- Interest is on to detect the semantically redundant paraphrases: non-novelty

Evaluation System	**Description**	**Precision**	**Recall**	**F-measure**	**Accuracy**
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	Set Difference+LR (Zhang et al., 2002)	0.71	0.52	0.60	64.75%
Novelty Measure 2	Geometric Distance+LR (Zhang et al., 2002)	0.69	0.75	0.71	70.23%
Novelty Measure 3	Language Model (KLD) +LR (Zhang et al., 2002)	0.74	0.77	0.75	74.34%
Novelty Measure 4	IDF+LR (Karkali et al., 2013)	0.65	0.55	0.59	61.72%
Proposed Approach	**RDV-CNN**	**0.75**	**0.84**	**0.80**	**78.02%**

Novelty Goes Deep: A Deep Neural Approach Towards Document Level Novelty Detection by Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, Sameer Chivukula and Georgios Tsatsaronis accepted as a full paper in the 27th International Conference on Computational Linguistics (COLING 2018) held at Santa Fe, New-Mexico, USA (CORE rank A/H-Index: 43).

Results on TAP-DLND 1.0

Novelty Goes Deep: A Deep Neural Approach Towards Document Level Novelty Detection by Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, Sameer Chivukula and Georgios Tsatsaronis accepted as a full paper in the 27th International Conference on Computational Linguistics (COLING 2018) held at Santa Fe, New-Mexico, USA (CORE rank A/H-Index: 43).

Is Your Document Novel? Let Attention Guide You. An Attention-Based Model For Document Level Novelty Detection by Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal and Pushpak Bhattacharyya provisionally accepted in Natural Language Engineering (NLE) Journal by Cambridge University Press

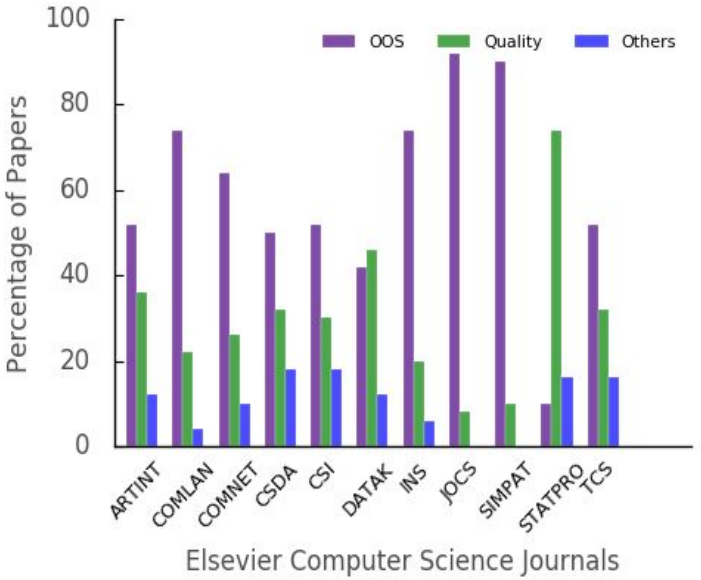
Evaluation System	**Description**	**P (N)**	**R (N)**	**F1 (N)**	**P (NN)**	**R (NN)**	**F1 (NN)**	**A (%)**
Baseline 1 (without SNLI pre-training)	Paragraph Vector+LR	0.75	0.75	0.75	0.69	0.69	0.69	72.81
Baseline 2 (without RDV-CNN)	BiLSTM+MLP	0.78	0.84	0.80	0.78	0.71	0.74	78.57
Novelty Measure 1	Set Difference+LR (Zhang et al., 2002)	0.74	0.71	0.72	0.72	0.74	0.73	73.21
Novelty Measure 1	Geometric Distance+LR (Zhang et al., 2002)	0.65	0.84	0.73	0.84	0.55	0.66	69.84
Novelty Measure 1	LM:(KLD)+LR (Zhang et al., 2002)	0.73	0.74	0.74	0.74	0.72	0.73	73.62
Novelty Measure 1	IDF+LR (Karkali et al., 2013)	0.52	0.92	0.66	0.66	0.16	0.25	54.26
(Ghosal et al., 2018)	Supervised Method (Feature-Based)	0.77	0.82	0.79	0.80	0.76	0.78	79.27
Proposed Approach	**RDV-CNN**	**0.86**	**0.87**	**0.86**	**0.84**	**0.83**	**0.83**	**84.53**
Proposed Approach	**Decomposable Attention Based**	0.85	0.85	0.85	**0.89**	**0.89**	**0.89**	**87.4**

Editorial Pre-Screening (Desk Rejection)

- Survey of ~7000
  - ACCEPTED (ACC)
  - DESK-REJECTED (DREJ)
  - REJECTED-AFTER-REVIEW (RAR)

  papers from 11 Elsevier Computer Science journals

- Study of corresponding *Author-Editor-Reviewer Interactions*
- Major Factors (DREJ):
  - **Appropriateness/Scope**
  - **Quality**
  - **Novelty**
  - Template Inconsistencies
  - Spelling, Language and Grammar



Investigating Impact Features in Editorial Pre-Screening of Research Papers by **Tirthankar Ghosal**, Rajeev Verma, Asif Ekbal, Sriparna Saha and Pushpak Bhattacharyya accepted as a Poster paper in the 18th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2018** held at Fort Worth, Texas, US from June 3-6, 2018 (CORE rank A*)

Exploring the Implications of Artificial Intelligence in Various Aspects of Scholarly Peer Review by **Tirthankar Ghosal** published at Doctoral Consortium of the 18th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2018** held at Fort Worth, Texas, US from June 3-6, 2018 (CORE rank A*), IEEE-TCDL

Task 2: Scope Detection

- AI assistance to editorial decisions
- Selection of an appropriate journal for a prospective manuscript
- Statistics reveal that about 25 - 50% of Desk rejections accounts for the article not being within the scope of the journal
- Considerable time is wasted in management jobs: Peer Review; Reduce the first turn-around time
- Current work : a machine learning based automated system to determine whether a submitted article falls into the scope of the journal

Investigating Impact Features in Editorial Pre-Screening of Research Papers by **Tirthankar Ghosal**, Rajeev Verma, Asif Ekbal, Sriparna Saha and Pushpak Bhattacharyya accepted as a Poster paper in the 18th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2018** held at Fort Worth, Texas, US from June 3-6, 2018 (CORE rank A*)

Exploring the Implications of Artificial Intelligence in Various Aspects of Scholarly Peer Review by **Tirthankar Ghosal** published at Doctoral Consortium of the 18th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2018** held at Fort Worth, Texas, US from June 3-6, 2018 (CORE rank A*), IEEE-TCDL

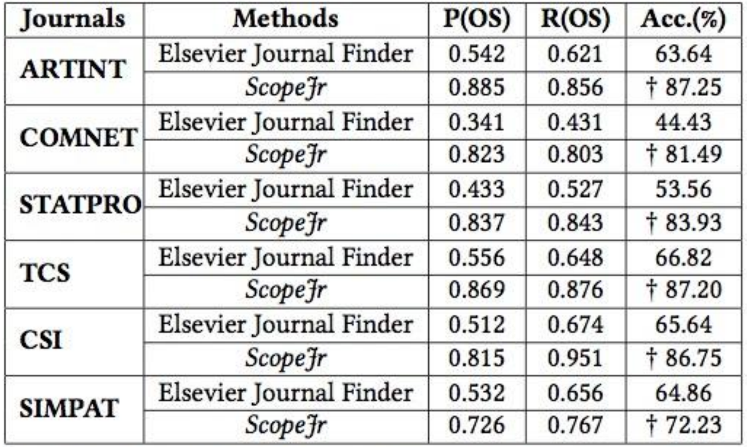
Scope Detection of Research Articles

- Having a universal definition of scope is hard. It varies across journals.
- Qualitatively we focus on the topics covered or domain of operation of the journal
- Problem framed as a binary classification problem in machine learning (IS : inscope / OS : outscope)
- Rejected data provided by a reputed publishing house (Elsevier)
- Extensive Feature Engineering to churn out reasons for Desk rejection from real data
- ***Start up idea :***
  - ***"Bibliographic information plays a major role in determining scope of a scholarly article : If an article belongs to a certain domain then it is seen that majority of its references falls into that domain"***
  - ***When a certain portion of a scientific article cites an in-domain reference, the scope of that portion is influenced by the domain of that reference. That is to say, the latent domain of the cited reference exerts local influence on that portion of the scientific article.***

Is the Paper Within Scope? Are You Fishing in the Right Pond? by **Tirthankar Ghosal**, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya and Ravi Sonam published as short paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)

Features-based (Approach-I)

- Bibliographic Features (of a candidate article Y)
  - Title Score
    - $T_Y = \sum V(T_i)$ (i=1 to m ; m is the number of references in Y)
  - Conference Score
    - $C_Y = \sum V(C_i)$ (i=1 to m ; m is the number of conference references in Y)
  - Journal Score
    - $J_Y = \sum V(J_i)$ (i=1 to m ; m is the number of journal references in Y)
- Author Journal Publication Frequency
- Keyword Overlap Score
- Distance from Cluster Boundary of similar articles



Journals	Methods	P(OS)	R(OS)	Acc.(%)
ARTINT	Elsevier Journal Finder	0.542	0.621	63.64
	ScopeJr	0.885	0.856	† 87.25
COMNET	Elsevier Journal Finder	0.341	0.431	44.43
	ScopeJr	0.823	0.803	† 81.49
STATPRO	Elsevier Journal Finder	0.433	0.527	53.56
	ScopeJr	0.837	0.843	† 83.93
TCS	Elsevier Journal Finder	0.556	0.648	66.82
	ScopeJr	0.869	0.876	† 87.20
CSI	Elsevier Journal Finder	0.512	0.674	65.64
	ScopeJr	0.815	0.951	† 86.75
SIMPAT	Elsevier Journal Finder	0.532	0.656	64.86
	ScopeJr	0.726	0.767	† 72.23

Table 4: Scope-Check figures for *out-of-scope* (OS) class across 6 journals, *P* → *Precision*, *R* → *Recall*. The Accuracy values (†) for *ScopeJr* are statistically significant over EJF performance (two-tailed t-test, p<0.05)

Is the Paper Within Scope? Are You Fishing in the Right Pond? by **Tirthankar Ghosal**, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya and Ravi Sonam published as short paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)

Dataset Description

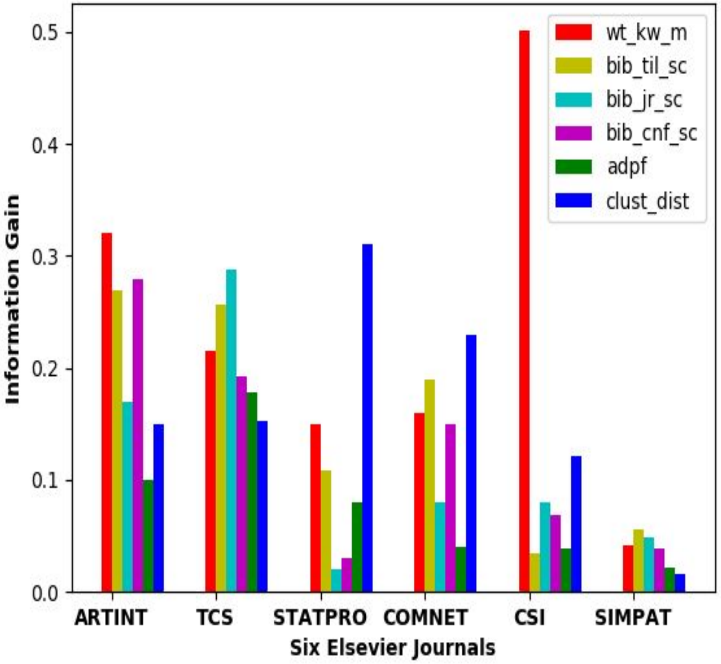
Dataset- I [Elsevier Computer Science Journals]

ARTINT, COMNET, JNCA, CSI, SIMPAT and STATPRO.

Dataset - II [Open Access]

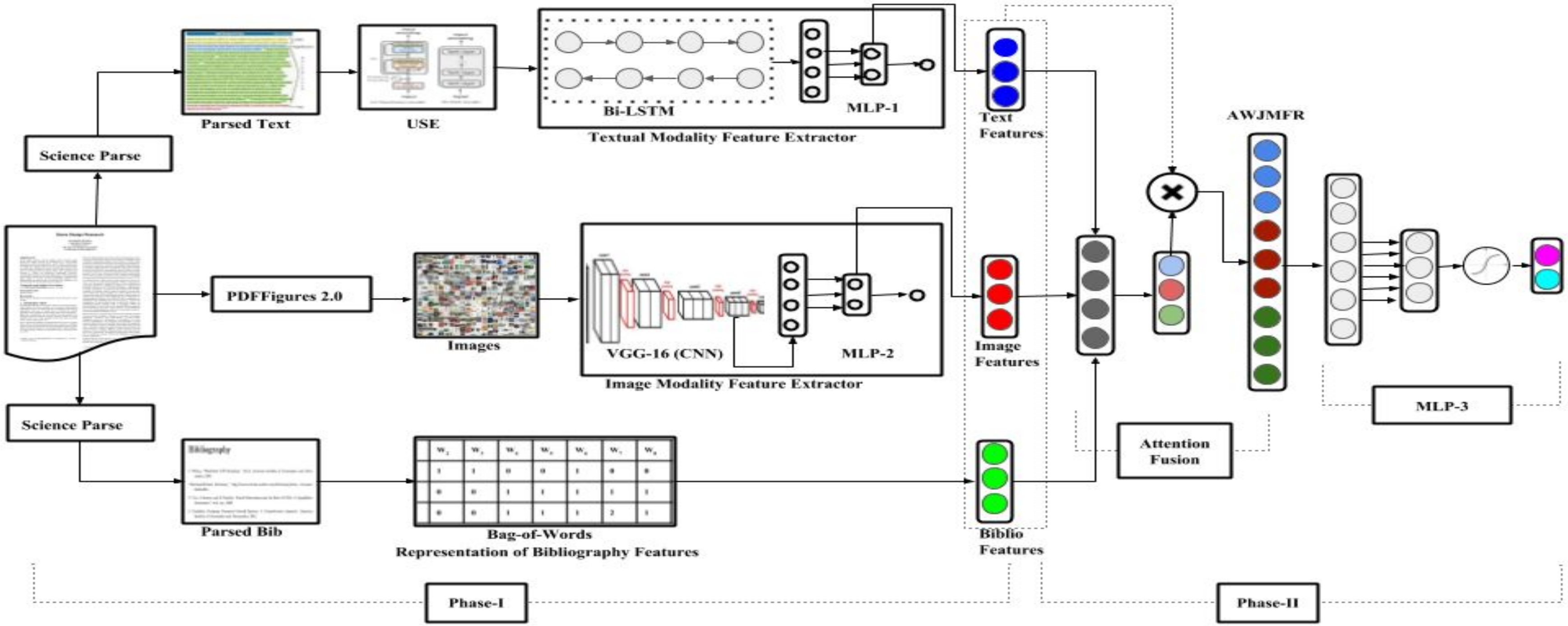
For AI/ML: ICLR, AAAI, IJCAI and NeurIPS (~ 7600 papers)
For NLP: ACL, NAACL, COLING, and CoNLL (~6700 papers)

For CV: CVPR, ECCV, and ICCV (~6400 papers)



Is the Paper Within Scope? Are You Fishing in the Right Pond? by **Tirthankar Ghosal**, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya and Ravi Sonam published as short paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)

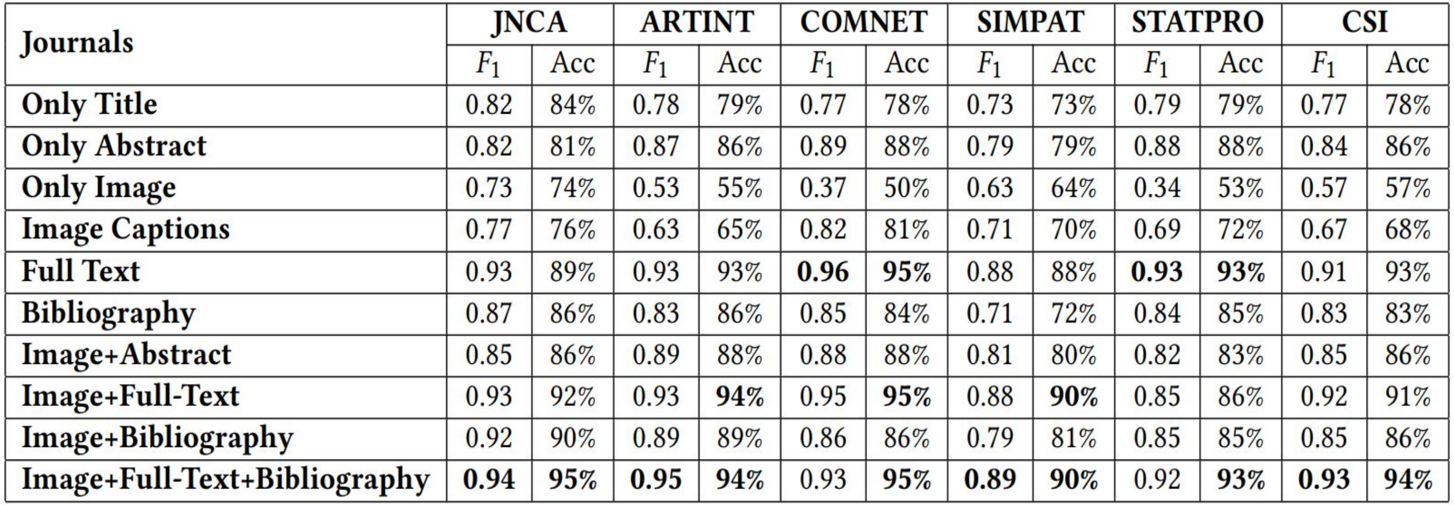
Approach II: Proposed Multimodal Deep Architecture



	w_1	w_2	w_3	w_4	w_5	w_6	w_7
	1	1	0	0	1	0	0
	0	0	1	1	1	1	1
	0	0	1	1	1	2	1

A Deep Multimodal Investigation To Determine the Appropriateness of a Scholarly Submission by **Tirthankar Ghosal**, Ashish Raj, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya accepted as full paper paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** to be held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)

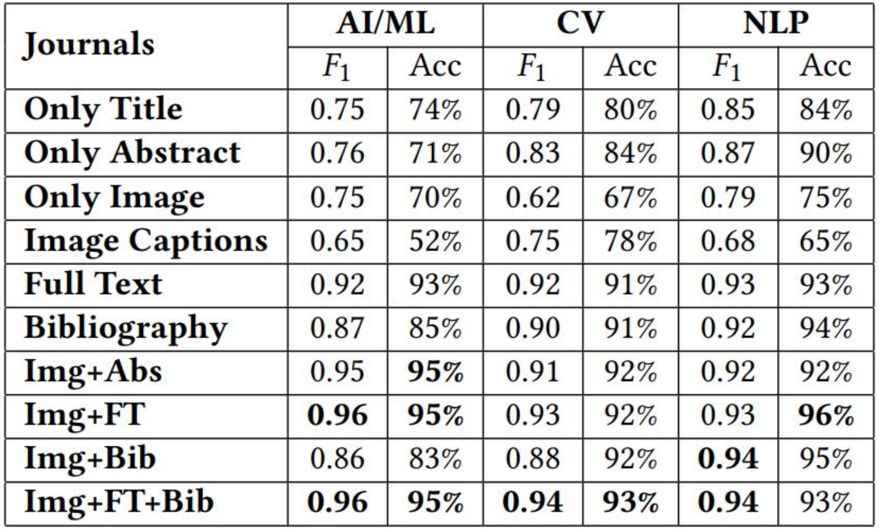
Results on Dataset-I (Journals)



Journals	JNCA		ARTINT		COMNET		SIMPAT		STATPRO		CSI	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
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%	**0.96**	**95%**	0.88	88%	**0.93**	**93%**	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	**94%**	0.95	**95%**	0.88	**90%**	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+Bibliography	**0.94**	**95%**	**0.95**	**94%**	0.93	**95%**	**0.89**	**90%**	0.92	**93%**	**0.93**	**94%**

A Deep Multimodal Investigation To Determine the Appropriateness of a Scholarly Submission by **Tirthankar Ghosal**, Ashish Raj, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya accepted as full paper paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** to be held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)

Results on Dataset-II (AI Conferences)



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

- ❖ How could we ascertain scope in similar domain?
- ❖ Towards an AI-powered recommender considering all aspects of a paper

A Deep Multimodal Investigation To Determine the Appropriateness of a Scholarly Submission by **Tirthankar Ghosal**, Ashish Raj, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya accepted as full paper paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** to be held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)

A Multiview Clustering Approach To Identify Out-of-Scope Submissions in Peer Review

Tirthankar Ghosal, Debomit Dey, Abhik Dutta, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya
Indian Institute of Technology Patna, Indian Institute of Engineering Science and Technology Shibpur
tirthankar.pcs16@iitp.ac.in

Introduction

- Peer Review is still the benchmark of research validation.
- A good number of submissions are desk-rejected for being not within the scope of the venue (25-30%) [2,5].
- Our current work tends to assist the editors to efficiently locate out-of-scope submissions.
- Our proposed semi-supervised approach is efficient and requires less training data to isolate out-of-scope submissions.

Problem Definition

- The problem is simple: To label a research article as out-of-scope if it does not fall within the scope of an intended journal (venue).
- The domain of operation or scope of a journal is defined by its past accepted articles.
- 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.

Solution Strategy

- 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.
- We apply a multi objective clustering (AMOSA) algorithm on each of the three views to generate the consensus partitions.
- Finally we apply K-Medoids to separate the unlabelled input data into In-Scope and Out-of-Scope clusters.

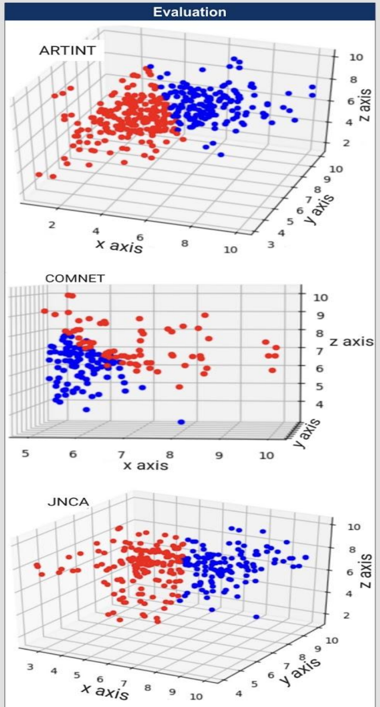
Dataset

- 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 library for information extraction from full-text and bibliography.

Methodology

- We represent the paper with the following views: Semantic, Lexical, and Bibliography. The core idea is to look at the paper through different perspectives where:
  1. Lexical view corresponds to the surface form of the text
  2. Semantic view would delve into the meaning representation of the full-text
  3. Bibliographic view would take into account the type of citations the paper contain. Citations are a valuable indication to the domain of operation of a scientific article [1].
- We use feature representations from all of these three views and remove stop words/irrelevant words from the scholarly texts.
- Semantic View: We use word2vec [3] concatenation on extracted keywords and entities to generate the semantic document representation of a research article.
- Lexical View: We adopt similar approach and use term frequency-inverse document frequency (tf-idf) as the lexical document representation.
- 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 [4] considering these three views to generate the consensus document clusters.
- 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.

Evaluation



- Figure 1 shows the efficiency of our approach tested on the three different journals.
- Multi-view Clustering of In-Scope and Out-of-Scope articles
- X-axis: Semantic View,Y-axis: Lexical View, Z-axis: Bibliography View

Results

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

Table 1: Cluster Prediction (*In-Scope* or *Out-Scope*) Results on the 3 journals,†→Baselines

Discussion

- Table 1 shows our results for the predicted cluster labels against the actual labels.
- We find that the Bibliography view is the most effective one [1].
- 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.

Conclusion and Future Work

- Here we explore multiview clustering to identify the appropriateness of an article to a journal.
- With little supervision, our method shows promise to address the continuously evolving peer review landscape.
- We intend to investigate our approach across more journals further and include the images present in the research articles as another view.

Acknowledgement

References

[1] Tirthankar Ghosal, Ravi Sonam, Sriparna Saha, Asif Ekbal, and Pushpak Bhattacharyya. 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) (7-12). European Language Resources Association (ELRA), Paris, France.

[2] Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018. 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. 333–334. https://doi.org/10.1145/3197026.3203910

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Je Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In Advances in Neural Information Processing Systems. 3111–3119.

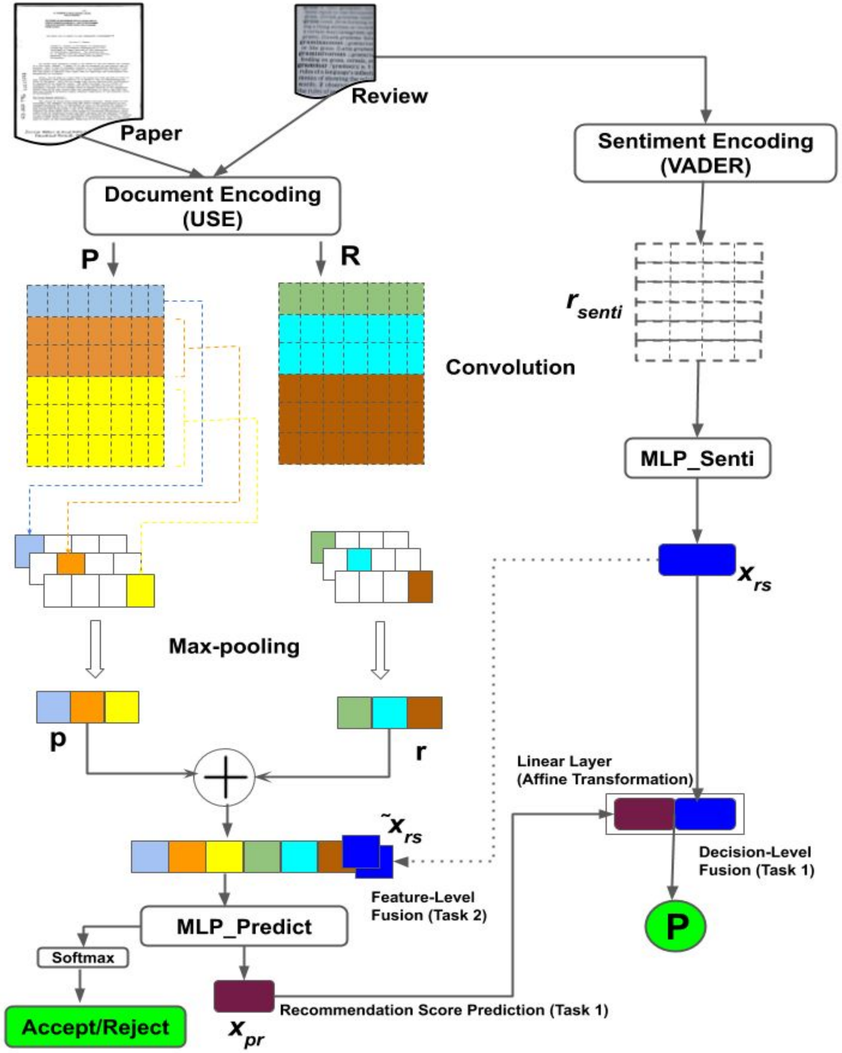
[4] Sriparna Saha, Sayantan Mitra, and Stefan Kramer. 2018. Exploring Multiobjective Optimization for Multiview Clustering. ACM Transactions on Knowledge Discovery from Data (TKDD) 12, 4 (2018), 44.

[5] Susan Trumbore, Mary-Elena Carr, and Sara Mikaloff-Fletcher. 2015. Criteria for Rejection of Papers Without Review. Global Biogeochemical Cycles 29, 8 (2015), 1123–1123.

Augment Sentiment to Peer Review Texts to Predict Outcome

- To predict the recommendation score and final decision from the interaction of reviews, paper, and reviewer sentiment
- ICLR 2017, 2018, 2019 papers
- Missing: Not all reviews are significant, interplay and correspondence between reviews and paper

DeepSentiPeer: Harnessing Sentiment in Review Texts To Recommend Peer Review Decisions by **Tirthankar Ghosal,** Rajeev Verma, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 57th Annual Meeting of the Association for Computational Linguistics (ACL) to be held at Florence (Italy) from July 28th to August 2nd, 2019 (CORE Rank A*, H5 Index: 106).



Paper
Review
Sentiment Encoding (VADER)
Document Encoding (USE)
P
R
r_{senti}
Convolution
MLP_Senti
x_{rs}
Max-pooling
p
r
Linear Layer (Affine Transformation)
$\tilde{x}_{rs}$
Decision-Level Fusion (Task 1)
Feature-Level Fusion (Task 2)
MLP_Predict
Softmax
P
Accept/Reject
x_{pr}
Recommendation Score Prediction (Task 1)

Publications

1. *TAP-DLND 1.0: A Corpus for Document Level Novelty Detection* by **Tirthankar Ghosal**, Amitra Salam, Swati Tiwari, Asif Ekbal and Pushpak Bhattacharyya published as a full paper at **LREC 2018** (H-Index: 43) held at Miyazaki, Japan
2. *Document Level Novelty Detection: Textual Entailment Lends a Helping Hand* by Tanik Saikh, **Tirthankar Ghosal**, Asif Ekbal and Pushpak Bhattacharyya published in the proceedings of **ICON 2017** at Jadavpur University, Kolkata, India
3. Sentiment analysis on (Bengali horoscope) corpus by **Tirthankar Ghosal,** Sajal Kanti Das and Saprativa Bhattacharjee published as a full paper in **IEEE INDICON 2015** held at Jamia Milia Islamia University, New Delhi, India
4. *Can your paper evade the editors axe? Towards an AI assisted peer review system* by **Tirthankar Ghosal**, Rajeev Verma, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya [https://arxiv.org/pdf/1802.01403.pdf]
5. *Investigating Impact Features in Editorial Pre-Screening of Research Papers* by **Tirthankar Ghosal**, Rajeev Verma, Asif Ekbal, Sriparna Saha and Pushpak Bhattacharyya accepted as a Poster paper in the 18th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2018** held at Fort Worth, Texas, US from June 3-6, 2018 (CORE rank A*)
6. *Exploring the Implications of Artificial Intelligence in Various Aspects of Scholarly Peer Review* by **Tirthankar Ghosal** accepted in Doctoral Consortium of the 18th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2018** held at Fort Worth, Texas, US from June 3-6, 2018 (CORE rank A*)
7. *Investigating Domain Features For Scope Detection and Classification of Scientific Articles* by **Tirthankar Ghosal**, Ravi Sonam, Sriparna Saha, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 7th International Workshop on Mining Scientific Publications **(WOSP 2018)** held in conjunction with LREC 2018 at Miyazaki, Japan from May 7-11, 2018.
8. *Novelty goes Deep: A Deep Neural Approach Towards Document Level Novelty Detection* by **Tirthankar Ghosal**, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, Sameer Chivukula and Georgios Tsatsaronis accepted as a full paper in the 27th International Conference on Computational Linguistics **(COLING 2018)** held at Santa Fe, New-Mexico, USA (CORE rank A/H-Index: 41).
9. *To Comprehend the New: On Measuring the Freshness of a Document* by **Tirthankar Ghosal**, Abhishek Shukla, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 37th International Joint Conference on Neural Networks **(IJCNN 2019)** to be held at Budapest, Hungary (CORE rank A/H-Index: 31).

Publications

10. *Is the Paper Within Scope? Are You Fishing in the Right Pond?* by **Tirthankar Ghosal**, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya and Ravi Sonam accepted as short paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** to be held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)
11. *A Deep Multimodal Investigation To Determine the Appropriateness of a Scholarly Submission* by **Tirthankar Ghosal**, Ashish Raj, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya accepted as full paper paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** to be held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)
12. *A Multiview Clustering Approach To Identify Out-of-Scope Submissions in Peer Review* by **Tirthankar Ghosal**, Debomit Dey, Avik Dutta, Asif Ekbal, Sriparna Saha and Pushpak Bhattacharyya accepted as poster paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** to be held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)
13. *Incorporating Full Text and Bibliographic Features to Improve Scholarly Journal Recommendation* by **Tirthankar Ghosal,** Ananya Chakraborty, Ravi Sonam, Asif Ekbal, Sriparna Saha and Pushpak Bhattacharyya accepted as poster paper in the 19th ACM/IEEE Joint Conference on Digital Libraries **(JCDL) 2019** to be held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)
14. *Is Your Document Novel? Let Attention Guide You. An Attention-Based Model For Document Level Novelty Detection* by **Tirthankar Ghosal**, Vignesh Edithal, Asif Ekbal and Pushpak Bhattacharyya provisionally accepted in Natural Language Engineering (NLE) Journal by Cambridge University Press
15. *A Sentiment Augmented Deep Architecture to Predict Peer Review Outcomes* by **Tirthankar Ghosal**, Rajeev Verma, Asif Ekbal and Pushpak Bhattacharyya accepted as poster paper in the 19th ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2019 to be held at University of Illinois Urbana-Champaign, US from June 2-5, 2019 (CORE rank A*)
16. *DeepSentiPeer: Harnessing Sentiment in Review Texts To Recommend Peer Review Decisions* by **Tirthankar Ghosal,** Rajeev Verma, Asif Ekbal and Pushpak Bhattacharyya accepted as a full paper in the 57th Annual Meeting of the Association for Computational Linguistics (ACL) to be held at Florence (Italy) from July 28th to August 2nd, 2019 (CORE Rank A*, H5 Index: 106).

Communities to explore:

- JCDL
- ASIS&T
- FORCE
- *CL
- SSP
- SIGIR
- KDD
- Peer Review Week
- Research on Research

References

[1] James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM.

[2] James Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in tdt is hard. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 374–381. ACM.

[3] James Allan, Courtney Wade, and Alvaro Bolivar. 2003. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 314–321. ACM.

[4] Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. 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*, pages 632–642.

References

[6] Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase Acquisition via Crowdsourcing and Machine Learning. *Transactions on Intelligent Systems and Technology (ACM TIST)*, 4(3):43:1–43:21, June.

[7] Jaime Carbonell and Jade Goldstein. 1998. 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*, pages 335–336. ACM.

[8] Praveen Chandar and Ben Carterette. 2013. Preference based evaluation measures for novelty and diversity. In *SIGIR*.

[9] Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the trec 2004 terabyte track. In *TREC*, volume 4, page 74.

[10] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Bu ̈ttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *SIGIR*.

[11] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity. In *WSDM*.

References

[12] Kevyn Collins-Thompson, Paul Ogilvie, Yi Zhang, and Jamie Callan. 2002. Information filtering, novelty detection, and named-page finding. In *TREC*.

[13] Ronan Collobert,Jason Weston,Leon Bottou,Michael Karlen,Koray Kavukcuoglu,and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493– 2537.

[14] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

[15] Tirthankar Dasgupta and Lipika Dey. 2016. Automatic scoring for innovativeness of textual ideas. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

[16] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. 2004. News Junkie: providing personalized news feeds via analysis of information novelty. In *Proceedings of the 13th international conference on World Wide Web*, pages 482–490. ACM.

[17] Tirthankar Ghosal, Amitra Salam, Swati Tiwary, Asif Ekbal, and Pushpak Bhattacharyya. 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.*

References

[18] Donna Harman. 2002. Overview of the TREC 2002 novelty track. In *Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002*.

[19] Margarita Karkali,Francois Rousseau,Alexandros Ntoulas, and Michalis Vazirgiannis. 2013. Efficient online novelty detection in news streams. In *WISE (1)*, pages 57–71.

[20] Yoon Kim. 2014. 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*, pages 1746–1751.

[21] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

[22] Xiaoyan Li and W Bruce Croft. 2005. Novelty detection based on sentence level patterns. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 744–751. ACM.

[23] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

References

[24] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree- based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

[25] Liyun Ru, Le Zhao, Min Zhang, and Shaoping Ma. 2004. Improved feature selection and redundancy computing- thuir at trec 2004 novelty track. In *TREC*.

[26] Barry Schiffman and Kathleen R McKeown. 2005. Context and learning in novelty detection. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 716–723. Association for Computational Linguistics.

[27] Ian Soboroff and Donna Harman. 2003. Overview of the TREC 2003 novelty track. In *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, pages 38–53.

[28] Ian Soboroff and Donna Harman. 2005. Novelty detection: the trec experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112. Association for Computational Linguistics.

References

[29] Tomoe Tomiyama, Kosuke Karoji, Takeshi Kondo, Yuichi Kakuta, Tomohiro Takagi, Akiko Aizawa, and Teruhito Kanazawa. 2004. Meiji university web, novelty and genomic track experiments. In *TREC*.

[30] Flora S Tsai and Yi Zhang. 2011. D2s: Document-to-sentence framework for novelty detection. *Knowledge and information systems*, 29(2):419–433.

[31] Arnout Verheij, Allard Kleijn, Flavius Frasincar, and Frederik Hogenboom. 2012. A comparison study for novelty control mechanisms applied to web news stories. In *Web Intelligence and Intelligent Agent Technology (WI- IAT), 2012 IEEE/WIC/ACM International Conferences on*, volume 1, pages 431–436. IEEE.

[32] Charles L Wayne. 1997. Topic detection and tracking (tdt). In *Workshop held at the University of Maryland on*, volume 27, page 28. Citeseer.

[33] Yiming Yang, Tom Pierce, and Jaime Carbonell. 1998. 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*, pages 28–36. ACM.

[34] Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 688–693. ACM.

References

[35] Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

[36] Yi Zhang and Flora S Tsai. 2009. Combining named entities and tags for novel sentence detection. In *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 30–34. ACM.

[37] Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88. ACM.

[38] Min Zhang, Chuan Lin, Yiqun Liu, Leo Zhao, and Shaoping Ma. 2003a. THUIR at TREC 2003: Novelty, robust and web. In *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, pages 556–567.

[39] Min Zhang, Ruihua Song, Chuan Lin, Shaoping Ma, Zhe Jiang, Yijiang Jin, Yiqun Liu, Le Zhao, and S Ma. 2003b. Expansion-based technologies in finding relevant and new information: Thu trec 2002: Novelty track experiments. *NIST SPECIAL PUBLICATION SP*, (251):586–590.

Acknowledgements

भारतीय प्रौद्योगिकी संस्थान पटना

INDIAN INSTITUTE OF TECHNOLOGY PATNA

सत्यमेव जयते

Ministry of Electronics &
Information Technology
Government of India

Visvesvaraya PhD Scheme for Electronics and IT

OAK RIDGE
National Laboratory

NON SOLUS

ELSEVIER

SIGIR
Special Interest Group
on Information Retrieval

IARCS

ORISE

Microsoft

FORCE11
The Future of Research Communications and e-Scholarship

acm

Crossref

ACL =
Annual Meeting of the Association for Computational Linguistics

a = a + c + l

THANK YOU !!

QUESTIONS ??

tirthankar.pcs16@iitp.ac.in
tirthankar.slg@gmail.com

LinkedIn	: https://www.linkedin.com/in/tirthankar-ghosal-ai/
Twitter	: @TirthankarSlg
Web	: https://tirthankarslg.wixsite.com/ainlpmldl