Sentiment Analysis in Indian languages

Md Shad Akhtar
Research Scholar
Department of Computer Science & Engineering
IIT Patna

Supervised by: Dr Asif Ekbal
Outline

- Definition & Motivation
- Challenges
- Recent Works
  - Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification
  - Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets
  - SAIL for Tweets
    - AMRITA-CEN@SAIL2015: Sentiment Analysis in Indian Languages
    - IIT-TUDA: System for Sentiment Analysis in Indian Languages Using Lexical Acquisition
  - Aspect Based Sentiment Analysis
    - Aspect based Sentiment Analysis in Hindi: Resource Creation and Evaluation
    - Aspect Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi
  - A Deep Learning Architecture for Multi-domain Sentiment Analysis
DEFINITION & MOTIVATION
WHAT IS SENTIMENT ANALYSIS?

- भावनाओं का विश्लेषण (*bhaavanaon ka vishleshan*)
- Sentiment analysis aims to identify the orientation of opinion in a piece of text.
**What is Sentiment Analysis?**

- Few examples:

<table>
<thead>
<tr>
<th></th>
<th>Devanagari</th>
<th>Transliterated</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Text</td>
<td>यह मूवी अच्छी नहीं है।</td>
<td>yah moovee Achchhee naheeN hai</td>
<td>😞</td>
</tr>
<tr>
<td></td>
<td>कोई भी मूवी इस से अच्छी नहीं हो सकती।</td>
<td>koEE bhee moovee Is se Achchhee naheeN ho sakatee</td>
<td>😊</td>
</tr>
<tr>
<td></td>
<td>इस मोबाइल का कैमरा अच्छा है।</td>
<td>Is mobaaIl kaa kaimaraa Achchhaa hai</td>
<td>😊</td>
</tr>
</tbody>
</table>
“What people think?”

What others think has always been an important piece of information.

मैं कौन सा मोबाइल खरीदूं?

“Which mobile should I buy?”
“So whom shall I ask?”

Pre Web

- Friends and relatives
- Acquaintances
- Consumer Reports

Post Web

- Blogs (google blogs, livejournal)
- E-commerce sites (flipkart, amazon, ebay)
- Review sites (CNET, PC Magazine)
- Discussion forums (forums.macrumors.com)
- Friends and Relatives (occasionally)

“... यह मोबाइल बहुत अच्छा है। इसका कैमरा अच्छा काम करता है। ...” (yah mobaall bahut Achchhaa hai. Isakaa kaimaraa Achchhaa kaam karataa hai. ... )
“Whoala! I have the reviews I need”

Now that I have “too much” information on one topic… I could easily form my opinion and make decisions…

Is this true?
Not Quite

• Searching for reviews may be difficult
  – Can you search for opinions as conveniently as general web search?
    eg: Is it easy to search for “IPhone vs Samsung Phone”?

• Overwhelming amounts of information on one topic
  – Difficult to analyze each and every review
  – Reviews are expressed in different ways
    • “सैमसंग का फ़ोन बहुत ही बेकार है।” (samsung kaa phone bahut hee bekaara hai.. )
    • “इस फ़ोन पर मेरे पैसे बरबाद हो गए।” (Is phone parra mere paise barabaad ho gaE.. )
    • “सैमसंग से अच्छा मैं Iphone खरीद लेता।” (samsung se Achchhaa maiN Iphone khareed letaa..)
“Let me look at reviews on one site only…”

Problems?

- Biased views
  - all reviewers on one site may have the same opinion

- Fake reviews/Spam
  - people post good reviews about their own products OR services
  - some posts are plain spams
An example…

- Mr. X needs to buy a phone but he is not sure which one to choose. So, he went to flipkart.com and browse the reviews of a particular phone.

- **Scenario 1:**
  - Suppose there are 1000 reviews out of which 850 reviews are negative, 100 are positive and rest 50 are neutral.
  - Overall polarity of the phone? 😞
  - Questions?
    - Can he read all the reviews? (Very less chance)
    - What if all the 100 positive reviews are at the top?

- **Scenario 2:**
  - Suppose there are 1000 reviews out of which 480 reviews are positive, 420 are negative and rest 100 are neutral.
  - Overall polarity of the phone? 😊
  - Questions?
    - Can he read all the reviews? (Very less chance)
    - What if few of the reviews (e.g. 100) are fake?
Levels of Sentiment Analysis

1. Document level
2. Sentence level
3. Phrase level
4. Aspect level

Increasing level of information
Levels of Sentiment Analysis

• Document level – Sentiment of complete document

<table>
<thead>
<tr>
<th>Document 1</th>
<th>Document 2</th>
<th>Document n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence 1</td>
<td>Sentence 1</td>
<td>Sentence 1</td>
</tr>
<tr>
<td>Sentence 2</td>
<td>Sentence 2</td>
<td>Sentence 2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Sentence n</td>
<td>Sentence n</td>
<td>Sentence n</td>
</tr>
</tbody>
</table>

Positive  
Negative  
Positive
Levels of Sentiment Analysis

- **Sentence level** — Sentiment of each sentence

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence 1</td>
<td>Positive</td>
</tr>
<tr>
<td>Sentence 2</td>
<td>Negative</td>
</tr>
<tr>
<td>Sentence 3</td>
<td>Negative</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Sentence n</td>
<td>Positive</td>
</tr>
</tbody>
</table>
Levels of Sentiment Analysis

- Phrase level — Sentiment \( w.r.t. \) given phrase

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>w1 w2 w3 w4 w5 w6 ...</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence 2</td>
<td>w1 w2 w3 w4 w5 w6 ...</td>
<td>Negative</td>
</tr>
<tr>
<td>Sentence 3</td>
<td>w1 w2 w3 w4 w5 w6 ...</td>
<td>Negative</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Sentence n</td>
<td>w1 w2 w3 w4 w5 w6 ...</td>
<td>Positive</td>
</tr>
</tbody>
</table>
Levels of Sentiment Analysis

- **Aspect level** - Sentiment *w.r.t.* attribute of a product or service

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>w1 w2 <strong>w3</strong> w4 w5 w6 ...</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence 2</td>
<td><strong>w1</strong> w2 w3 w4 w5 w6 ...</td>
<td>Negative</td>
</tr>
<tr>
<td>Sentence 3</td>
<td>w1 <strong>w2 w3</strong> w4 w5 w6 ...</td>
<td>Negative</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Sentence n</td>
<td>w1 w2 w3 w4 <strong>w5</strong> w6 ...</td>
<td>Positive</td>
</tr>
</tbody>
</table>
Aspect Based Sentiment Analysis (ABSA)

- High level (Document or Sentence) sentiment analysis do not discover what exactly people liked and did not like!
- Opinion consists of a sentiment (positive, negative, neutral or conflict) and target of opinion.
- Opinion targets helps us to understand the sentiment analysis problem better.
- E.g:
  - इसकी बैटरी शानदार है, लेकिन कैमरा बहुत ही खराब है। (Isakee baiTaree shaanadaara hai, lekin kaimaraa bahut hee kharaab hai.. )
  - Positive about the battery but negative about the camera
Aspect Based Sentiment Analysis (ABSA)

- Four subtasks
  - Aspect Term Extraction (ATE) — Sequence labeling
    - Feature or attributes of a product or service
  - Aspect Term Sentiment (ATS) — Classification
  - Aspect Category Detection (ACD) — Multi-label classification
    - Generalization of aspect term
  - Aspect Category Sentiment (ACS) — Classification

<table>
<thead>
<tr>
<th>Subtasks</th>
<th>Review Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devanagari</td>
<td>“इसका हाउसिंग स्टेनलेस स्टील से निर्मित है इसलिए बहुत भारी है।”</td>
</tr>
<tr>
<td>Transliterated</td>
<td>“Isakaa haaUsiNg sTenaless sTeel se nirmit hai IsaliE bahut bhaaree hai.”</td>
</tr>
<tr>
<td>Translated</td>
<td>“Its housing is made up of stainless steel that why it is very heavy.”</td>
</tr>
<tr>
<td>ATE</td>
<td>हाउसिंग (haaUsiNg)</td>
</tr>
<tr>
<td></td>
<td>neutral</td>
</tr>
<tr>
<td>ATS</td>
<td>Design, Misc</td>
</tr>
<tr>
<td>ACD</td>
<td>neutral, negative</td>
</tr>
<tr>
<td>ACS</td>
<td></td>
</tr>
</tbody>
</table>
## Sentence Based v/s Aspect Based Sentiment Analysis

### Camcorder X

- The **zoom** is excellent, but the **LCD** is blurry.
- Great value for the **price**.
- Although the **display** is poor the **picture quality** is amazing.
- **Batteries** drain pretty quickly.
- I love this camera but for short **battery life** is definitely a pain.
- It is good camera for the **price**.
- ...

---

### Sentence based sentiment analysis

<table>
<thead>
<tr>
<th>Product</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camcorder X</td>
<td>3.1</td>
</tr>
</tbody>
</table>

### Aspect based sentiment analysis

<table>
<thead>
<tr>
<th>Aspect Term</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zoom</td>
<td>5</td>
</tr>
<tr>
<td>Price</td>
<td>4</td>
</tr>
<tr>
<td>Picture quality</td>
<td>4</td>
</tr>
<tr>
<td>Battery life</td>
<td>2</td>
</tr>
<tr>
<td>Screen</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
CHALLENGES
Challenges in Indian Languages

• Major challenges

1. Free word order: Most of the Indian languages (e.g. Hindi) follow free word order
   • कैमरा अच्छा है इस मोबाइल का। (kaimaraa Achchhaa hai Is mobaall kaa..)
   • अच्छा कैमरा है इस मोबाइल का। (Achchhaa kaimaraa hai Is mobaall kaa..)
   • इस मोबाइल का कैमरा अच्छा है। (Is mobaall kaa kaimaraa Achchhaa hai..)
Challenges in Indian Languages

- **Major challenges**

  2. Scarcity of various NLP tools and resources.
     - PoS tagger
     - Chunker
     - Dependency Parser
     - Sentiment Lexicons: A list of positive/negative words.

  3. Absence of benchmark datasets:
     - Quantity of reviews – few 100s
     - Quality of reviews – Translated reviews
RECENT WORKS
Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification

- Hindi Subjectivity lexicons
  - A subjective lexicons generated through WordNet seed words expansion techniques.
- Google translated product reviews
- Accuracy
  - Baseline: 74.62%
  - Subjectivity lexicons: 79.03%

Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets

- Multidict synset ID is used as feature.

<table>
<thead>
<tr>
<th>Synset Identifier</th>
<th>Hindi</th>
<th>Marathi</th>
</tr>
</thead>
<tbody>
<tr>
<td>13104</td>
<td>अवकाश (avkasha)</td>
<td>सुद्धी (suTTe)</td>
</tr>
<tr>
<td></td>
<td>छुट्टी (chuTTe)</td>
<td>रुजा (ruh-Jaa)</td>
</tr>
</tbody>
</table>

An example entry (concept: holiday) in Multidict for Hindi and Marathi

- IITB movie review dataset
  - Hindi – 200 sentences; Marathi – 150 sentences

- Accuracy
  - Hindi: 65.64% Baseline 83.06% MultiDict
  - Marathi: 86.53% Baseline 97.87% MultiDict

Balamurali, A. R., Joshi, A., and Bhattacharyya, P.; Cross-lingual sentiment analysis for Indian languages using linked wordnets. In COLING 2012
SENTIMENT ANALYSIS IN TWITTER
Sentiment Analysis in Twitter

• Problems
  • Unstructured data
  • Noisy text
    • Spelling variation – e.g. “grt, gr8, great” etc.
    • Elongation – e.g. “gooooooooooooddd”
  • Hashtags – e.g. #NotLikingIt
  • Usernames – e.g. @ImKohli
  • Images & links
  • Length restriction – max 140 chars
Sentiment Analysis in Indian Languages (SAIL) for Twitter - 2015

• Three languages
  • Tamil
  • Hindi
  • Bengali

• Datasets

<table>
<thead>
<tr>
<th>Language</th>
<th>Training data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Tamil</td>
<td>387</td>
<td>316</td>
</tr>
<tr>
<td>Hindi</td>
<td>168</td>
<td>545</td>
</tr>
<tr>
<td>Bengali</td>
<td>277</td>
<td>354</td>
</tr>
</tbody>
</table>

Patra, Braja Gopal and Das, Dipankar and Das, Amitava and Prasath, Rajendra; Shared task on Sentiment Analysis in Indian Languages (sail) Tweets-An Overview; MIKE. Springer - 2015
Sentiment Analysis in Indian Languages (SAIL) for Twitter - 2015

• AMRITA-CEN@SAIL2015: Sentiment Analysis in Indian Languages
  • Features used:
    • SentiWordNET
    • Binary features (#hashtags, @user, ?questionmark, !exclamation etc.)
  • Classifier
    • Naïve - Bayes
  • Evaluation
    • Tamil: 39.28%
    • Hindi: 55.67%
    • Bengali: 33.60%

Se, Shriya and Vinaya kumar, R. and Anand Kumar, M. and Soman, K. P.; AMRITA-CEN@SAIL2015: Sentiment Analysis in Indian Languages; MIKE. Springer- 2015
Sentiment Analysis in Indian Languages (SAIL) for Twitter - 2015

- IIT-TUDA: System for Sentiment Analysis in Indian Languages Using Lexical Acquisition
  - Lexical Acquisition: Words that occur in same context tend to have similar meanings.
    - Distributional Thesaurus (DT): An automatically computed resource that relates words according to their similarity

A. Kumar, S. Kohail, A. Ekbal and C. Biemann; IIT-TUDA: System for Sentiment Analysis in Indian Languages Using Lexical Acquisition; MIKE. Springer- 2015
Sentiment Analysis in Indian Languages (SAIL) for Twitter - 2015

• IIT-TUDA: System for Sentiment Analysis in Indian Languages Using Lexical Acquisition

• Lexical Acquisition: Words that occur in same context tend to have similar meanings.
  • Co-Occurences (CooC): A list of words that co-occur significantly with other words in a sentence.

A. Kumar, S. Kohail, A. Ekbal and C. Biemann; IIT-TUDA: System for Sentiment Analysis in Indian Languages Using Lexical Acquisition; MIKE. Springer- 2015
Aspect based sentiment analysis (ABSA)
Aspect Based Sentiment Analysis (ABSA) in Hindi: Resource Creation and Evaluation

• Resource creation for aspect term extraction and aspect term sentiment.

1. Data crawling
   • Crawled news, blogs, e-comm website.
   • Collected reviews across 12 domains
     – Mobile, Laptop, Tablet, Camera, Smart watches, Home Appliances, Head Phones, Speaker, Television, Mobile Apps, Travel and Movies.
   • Total 8000 reviews
Aspect Based Sentiment Analysis (ABSA) in Hindi: Resource Creation and Evaluation

• Resource creation for aspect term extraction and aspect term sentiment.

2. Data preprocessing
   • Removed irrelevant data.
   • Corrected obvious mistakes

<table>
<thead>
<tr>
<th>Original (Devanagari)</th>
<th>स्क्रीन का रिज़ोल्यूशन 1024 गुणा 600 है, जो काफी अच्छा है</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original (Transliterated)</td>
<td>skreen kaa riZolyooshan 1024 guNNaa 600 hai, jo kaaphee Achchh hai</td>
</tr>
<tr>
<td>Corrected</td>
<td>स्क्रीन का रिज़ोल्यूशन 1024 गुणा 600 है, जो काफी अच्छा है।</td>
</tr>
<tr>
<td>Corrected</td>
<td>skreen kaa riZolyooshan 1024 guNNaa 600 hai, jo kaaphee Achchhaa hai.</td>
</tr>
</tbody>
</table>

Akhtar, M. S., Ekbal, A. and Bhattacharyya, P.; Aspect Based Sentiment Analysis in Hindi: Resource Creation and Sentiment Classification; In proc. LREC 2016
Aspect Based Sentiment Analysis (ABSA) in Hindi: Resource Creation and Evaluation

• Resource creation for aspect term extraction and aspect term sentiment.

3. Data annotation
  • 5417 review sentences.
  • 3 human annotators.
  • Cohen’s Inter-rater agreement: 95.18%
Aspect Based Sentiment Analysis (ABSA) in Hindi: Resource Creation and Evaluation

• Evaluation
  – Features:
    • ATE – Word & context, N-grams, POS, Chunk, Prefix, Suffix etc.
    • ATS – Word & context, Word Bigram, Semantic Orientation Score (PMI)
  – Classifier:
    • ATE – Conditional Random Field (CRF)
    • ATS – Support Vector Machine (SVM)
  – Result:
    • ATE – 41.04% F-measure
    • ATS – 54.05% Accuracy
Aspect Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi

- Aspect category detection and aspect category sentiment

<table>
<thead>
<tr>
<th>Format</th>
<th>Review Text</th>
<th>Aspect Category</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devanagari</td>
<td>इसकी स्क्रीन 15.6 इंच की है।</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transliterated</td>
<td>Isakee skreen 15.6 INch kee hai.</td>
<td>Hardware</td>
<td>Neutral</td>
</tr>
<tr>
<td>Translated</td>
<td>It has 15.6 inch screen.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devanagari</td>
<td>यह बहुत महंगा है।</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transliterated</td>
<td>yah bahut mahanGaa hai.</td>
<td>Price</td>
<td>Negative</td>
</tr>
<tr>
<td>Translated</td>
<td>It is very costly.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Akhtar, M. S., Ekbal, A. and Bhattacharyya, P.; Aspect Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi; In proc. CICLING 2016
Aspect Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi

- Resource creation for ACD and ACS
- Similar to ATE and ATS
- Predefined set of aspect categories

<table>
<thead>
<tr>
<th>Domains</th>
<th>Aspect Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile apps</td>
<td>GUI, Ease of use, Price, Misc.</td>
</tr>
<tr>
<td>Travels</td>
<td>Scenery, Place, Reachability, Misc.</td>
</tr>
<tr>
<td>Movies</td>
<td>Story, Performance (Action/Direction etc.), Music, Misc.</td>
</tr>
</tbody>
</table>
Aspect Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi

• Evaluation

• Features:
  • ACD – N-grams, non contiguous N-grams, Character N-grams etc.
  • ACS – N-grams, non contiguous N-grams, POS, SO Score (PMI)

• Classifier:
  • ACD – Naïve Bayes, Decision Tree and SMO (MULAN framework)
  • ACS – Naïve Bayes, Decision Tree and SMO (WEKA framework)

• Result:
  • ACD – 46.46% (Electronics), 30.97% (Travels) and 56.53% (Mobile App), 64.27% (Movies)
  • ACS – 54.48% (Electronics), 65.20% (Travels) and 47.95% (Mobile App), 91.62% (Movies)

Akhtar, M. S., Ekbal, A. and Bhattacharyya, P.; Aspect Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi; In proc. CICLING 2016
Deep learning based sentiment analysis
A Deep Learning Architecture for Multi-domain Sentiment Analysis

• Convolutional Neural Network (CNN) based deep architecture
A Deep Learning Architecture for Multi-domain Sentiment Analysis

- Datasets
  - Product Reviews – Hindi (5217 reviews)
  - Twitter – Hindi (~1700 tweets)
    - Generic Tweets
  - Twitter – English (~10K tweets)
    - Generic Tweets
    - Sarcastic Tweets
A Deep Learning Architecture for Multi-domain Sentiment Analysis

- Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Twitter_H$</td>
<td>$Review_H$</td>
<td>$Twitter_{E\text{-Generic}}$</td>
<td>$Twitter_{E\text{-Sarcastic}}$</td>
</tr>
<tr>
<td>$B_{SVM}$</td>
<td>49.02</td>
<td>51.52</td>
<td>46.31</td>
<td>48.33</td>
</tr>
<tr>
<td>$CNN_W$</td>
<td>60.60</td>
<td>55.12</td>
<td>51.61</td>
<td>45.0</td>
</tr>
<tr>
<td>$CNN_{(W+X)}$</td>
<td>61.89</td>
<td>55.56</td>
<td>56.06</td>
<td>51.67</td>
</tr>
<tr>
<td>SAIL best system</td>
<td>55.60</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
THANK YOU!