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:

	Review Text	Polarity
Devanagari	यह मूवी अच्छी नहीं है।	
Transliterated	yah moovee Achchhee naheeN hai	
Devanagari	कोई भी मूवी इस से अच्छी नहीं हो सकती।	
Transliterated	koEE bhee moovee Is se Achchhee naheeN ho sakatee	
Devanagari	इस मोबाइल का कैमरा अच्छा है।	
Transliterated	Is mobaaIl kaa kaimaraa Achchhaa hai	

“What people think?”

What others think has always been an important piece of information

मैं कौन सा मोबाइल खरीदूं ?

maiN kaon saa mobaall khareedooN?

“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 para 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?

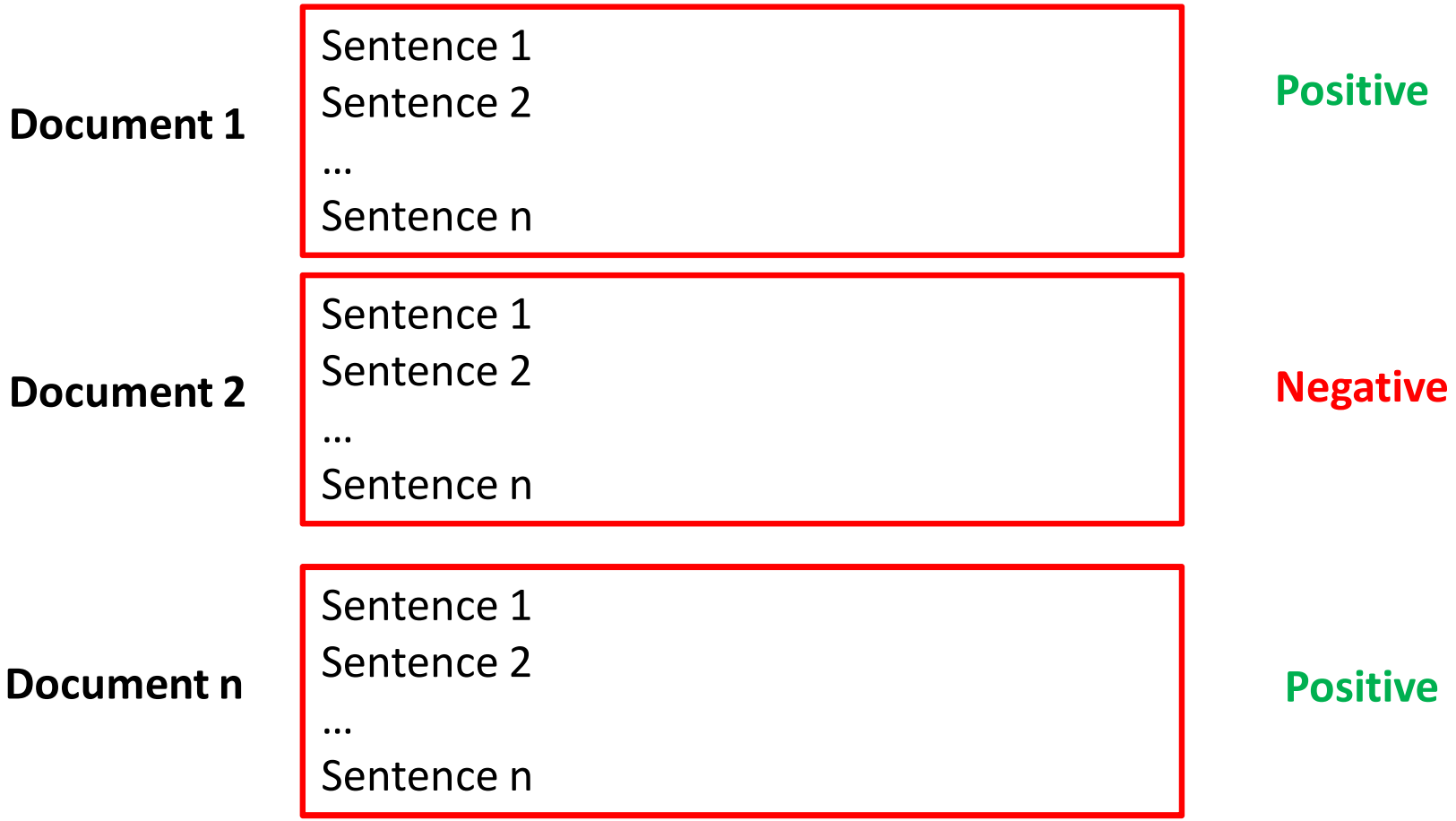
Levels of Sentiment Analysis

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



Levels of Sentiment Analysis

- Sentence level – Sentiment of each sentence

Sentence 1.	**Positive**
Sentence 2.	**Negative**
Sentence 3.	**Negative**
…	…
Sentence n	**Positive**

Levels of Sentiment Analysis

- Phrase level – Sentiment *w.r.t.* given phrase

Sentence 1	w1 w2 **w3 w4 w5** w6 …	**Positive**
Sentence 2	**w1 w2 w3 w4** w5 w6 …	**Negative**
Sentence 3	w1 **w2 w3** w4 w5 w6 …	**Negative**
…	…	**…**
Sentence n	w1 w2 **w3 w4 w5 w6** …	**Positive**

Levels of Sentiment Analysis

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

Sentence 1	w1 w2 **w3** w4 w5 w6 …	**Positive**
Sentence 2	**w1** w2 w3 w4 w5 w6 …	**Negative**
Sentence 3	w1 **w2 w3** w4 w5 w6 …	**Negative**
…	…	**…**
Sentence n	w1 w2 w3 w4 w5 w6 …	**Positive**

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 ख़राब 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**

Subtasks	Review Text	
	Devanagari	``इसका हाउसिंग स्टेनलेस स्टील से निर्मित है इसलिए बहुत भारी है।".
	Transliterated	``Isakaa haaUsiNg sTenales sTeel se nirmit hai IsaliE bahut bhaaree hai.".
	Translated	``Its housing is made up of stainless steel that why it is very heavy.".
ATE		हाउसिंग *(haaUsiNg)*
ATS		*neutral*
ACD		*Design, Misc*
ACS		*neutral, negative*

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.
- ..

Product	Rating
Camcorder X	3.1

Sentence based sentiment analysis

Aspect Term	Rating
Zoom	5
Price	4
Picture quality	4
Battery life	2
Screen	1
…	…

Aspect based sentiment analysis

CHALLENGES

Challenges in Indian Languages

- Major challenges
  1. Free word order: Most of the Indian lanaguages (e.g. Hindi) follow free word order
     - ***कैमरा अच्छा है इस मोबाइल का। (kaimaraa Achchhaa hai Is mobaaIl kaa..)***
     - ***अच्छा कैमरा है इस मोबाइल का। (Achchhaa kaimaraa hai Is mobaaIl kaa..)***
     - ***इस मोबाइल का कैमरा अच्छा है। (Is mobaaIl 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%**

Bakliwal, A., Arora, P., and Varma, V.; Hindi subjective lexicon: A lexical resource for Hindi polarity classification. *Int. J. Comput. Linguist. Appl.(IJCLA) - 2012*

Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets

- Multidict synset ID is used as feature.

Synset Identifier	Hindi	Marathi
13104	अवकाश (avkasha)	सुट्टी (suTTee)
	छुट्टी (chuTTee)	रजा (ruh-Jaa)

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. *“goooooooooooddd”*
  - 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

Language	Training data				Test data
	Positive	Negative	Neutral	Total	
Tamil	387	316	400	1103	560
Hindi	168	545	493	1222	467
Bengali	277	354	368	999	500

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

अतुलनीय (atulnIya)	अद्भुत (adabhuta)	महान (mahAna)	शानदार (shAnadAra)
तर्कसंगत (tarkasangata)	उचित (uchita)	सही (sahI)	गलत (galata)
धार्मिक (dhArmika)	सामाजिक (sAmAjika)	राजनीतिक (rAjanItika)	हिंदू (hindU)
ऊँची (UnchI)	ऊंची (UnchI)	लंबी (lambI)	छोटी (ChotI)

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-Occurrences (CooC): A list of words that co-occur significantly with other words in a sentence.

अतुलनीय (atulnIya)	भारतीय (bhAratIya)	अन्य (anya)	वर्ष (warSha)
तर्कसंगत (tarkasangata)	कहना (kahanA)	ज्यादा (jyadA)	काफी (kAphI)
धार्मिक (dhArmika)	परंपराओं (paramparAon)	अपितु (apitu)	संतों (santon)
ऊँची (UnchI)	इमारत (imArata)	जाति (jAti)	जगहों (jagahon)

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

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

Original (Devanagari)	स्क्रीन का रिज़ोल्यूशन 1024 गुणा 600 है, जो काफी अच्छ है
Original (Transliterated)	skreen kaa riZolyooshan 1024 guNNaa 600 hai , jo kaaphee Achchh hai
Corrected	स्क्रीन का रिज़ोल्यूशन 1024 गुणा 600 है , जो काफी अच्छा है।
Corrected	skreen kaa riZolyooshan 1024 guNNaa 600 hai , jo kaaphee Achchh**aa hai.**

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%**

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

- 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

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: Category Detection and Sentiment Classification for Hindi

- Aspect category detection and aspect category sentiment

Format	Review Text	Aspect Category	Sentiment
Devanagari	इसकी स्क्रीन 15.6 इंच की है।	Hardware	Neutral
Transliterated	Isakee skreen 15.6 INch kee hai.		
Translated	It has 15.6 inch screen.		
Devanagari	यह बहुत महंगा है।	Price	Negative
Transliterated	yah bahut mahaNgaa hai.		
Translated	It is very costly.		

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

Domains	Aspect Categories
Electronics (Laptops, Mobiles, Tablets, Cameras, Speakers, Smart watches, Headphones, Home appliances & Televisions)	Design, Software, Hardware, Ease of use, Price, Misc.
Mobile apps	GUI, Ease of use, Price, Misc.
Travels	Scenery, Place, Reachability, Misc.
Movies	Story, Performance (Action/Direction etc.), Music, Misc.

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

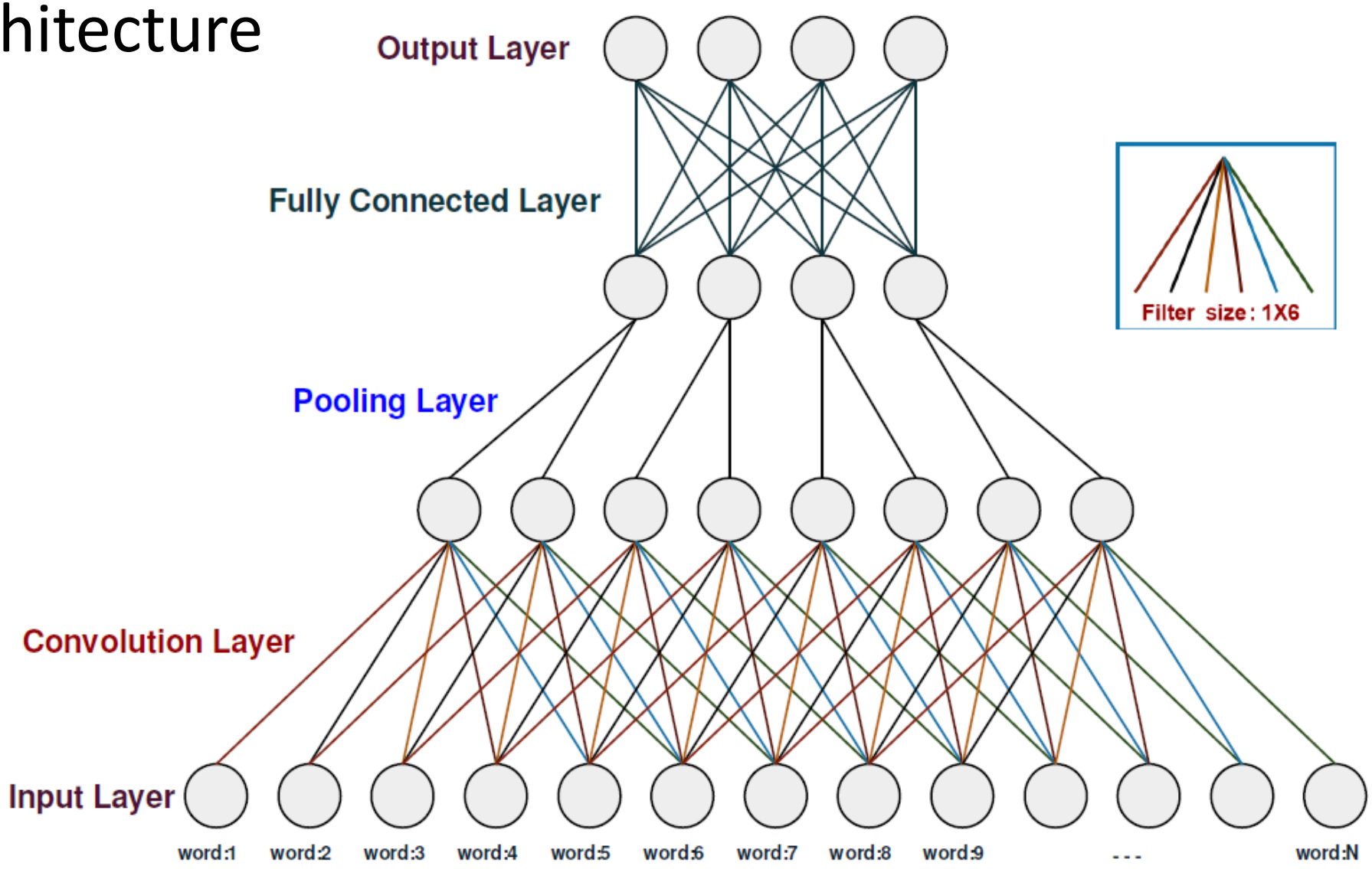
- Evaluation
  - Features:
    - ACD – N-grams, non contiguous N-grams, Character N-grams etc.
    - ACS – N-grams, non contiguous N-grams, POS, SO Score (PMI)
  - Classifier:
    - ACD – Naïve Bayes, Decision Tree and SMO (MULAN framework)
    - ACS – Naïve Bayes, Decision Tree and SMO (WEKA framework)
  - Result:
    - ACD – 46.46% (Electronics), 56.53% (Mobile App), 30.97% (Travels) and 64.27% (Movies)
    - ACS – 54.48% (Electronics), 47.95% (Mobile App), 65.20% (Travels) and 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

Methods	Accuracy			
	$Twitter_{H}$	$Review_{H}$	$Twitter_{E\text{-}Generic}$	$Twitter_{E\text{-}Sarcastic}$
B_{SVM}	49.02	51.52	46.31	48.33
CNN_{W}	60.60	55.12	51.61	45.0
$CNN_{(W+X)}$	61.89	55.56	56.06	51.67
SAIL best system	55.60	-	-	-

THANK YOU!