DEPARTMENT OF INFORMATION TECHNOLOGY
PILLAIS INSTITUTE OF INFORMATION TECHNOLOGY
ENGINEERING, MEDIA STUDIES AND RESEARCH
NEW PANVEL - 410 206
UNIVERSITY OF MUMBAI
Academic Year 2015-16

A

Stage One Dissertation Seminar On

'Sentiment Analysis of transliterated hindi and marathi script'

By

Mr. Mohammed Arshad Ansari (ME IT)

Under the Guidance of

Prof. Sharvari Govilkar

Outline

Introduction

- Presence of translited text in Literature
- Transliteration as vital source of sentiments
- Current state of affairs on the subject
- Why it matters now?

Outline

Motivation

- Reviews
- Discourse Analysis
- Feedback Analysis
- Other areas

Outline

Literature Survey

Paper	Author	Approach	Result	Limitation
Hindi subjective lexicon: A lexical resource for hindi poliarty classification [1]	Bakliwal, Akshat, Piyush Arora ad Vasudev Varma	Usage of sense based sentiment analysis and devlopment of HSWN	Resource for hindi	Doesn't work on transliteration of hindi text

Literature Survey

Paper	Author	Approach	Result	Limitation
A framework for sentiment analysis in hindi using HSWN [4]	Pooja Pandey , Sharvari Govilkar	Devnagiri sentiment analysis using HWSN and use of negation discourse analysis	Usage of negation and discourse analyse improves the result of polarity detection	Doesn't work with hindi transliteration

Literature Survey

Paper	Author	Approach	Result	Limitation
Automatic Detection of English words in Benglish text: A statistical approach [2]	Kundu, Bijoy and Chandra, Swarup	Statistical method developed which is language independent and can be used to detect any foreign language.	Accuracy upto nearly 72 percent.	There is a great scope of increasing the accuracy by improving methodology.

Literature Survey

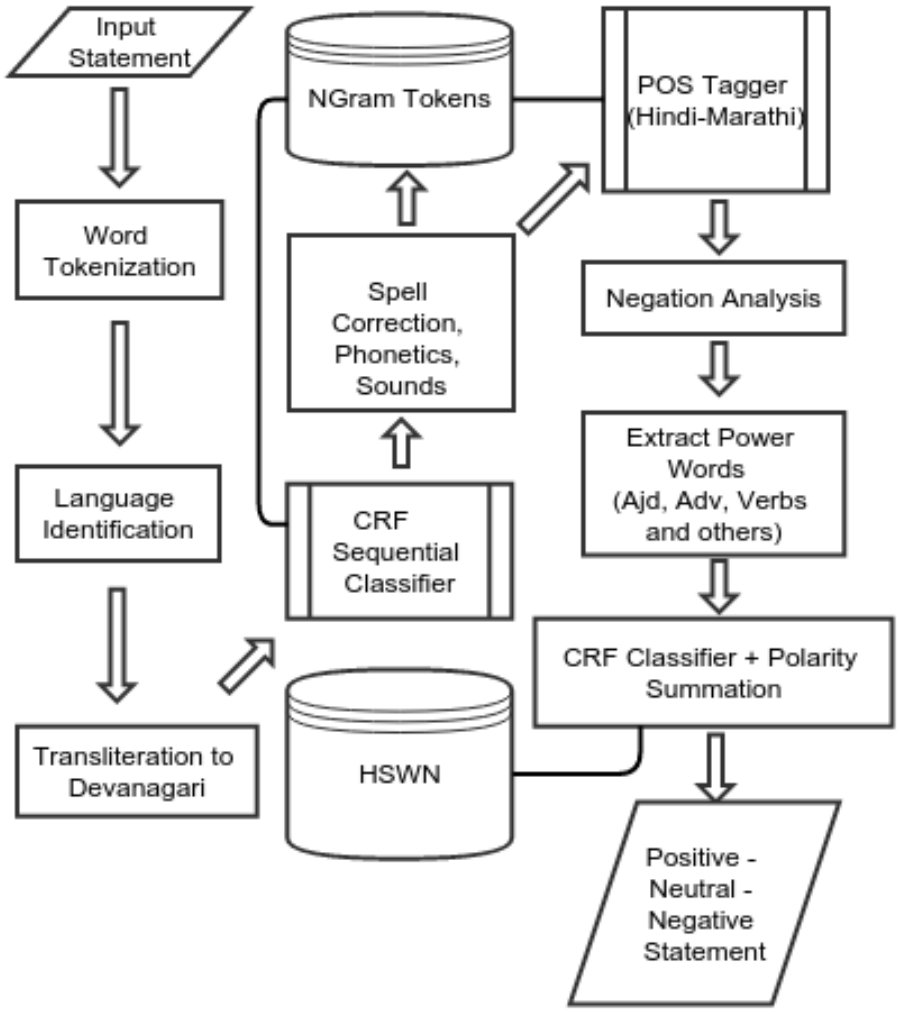
Paper	Author	Approach	Result	Limitation
Sentiment Analysis of Hindi Review based on negation and discourse relation [3]	Namita Mittal, Basant Agwarwal, Garvit Chouhan and Nitin Bania	Negation and discourse relation were identified and HSWN improvement carried out	Accuracy upto nearly 81 percent	Transl - iteration not covered. Scope for accuracy improvement.

Literature Survey

Paper	Author	Approach	Result	Limitation
Text Normalization of Code Mix and Sentiment Analysis [5]	PYKL Sriniva and Shashank Sharma	Language identification and transliteration to devnagiri script and then using HSWN for sentiment analysis.	Accuracy upto 85 precent	Negation and discourse analysis not being performed.

Outline

Proposed Approach: Flow diagram



Proposed Approach

1. Text Normalization
2. NLP and Sentiment Analysis

Proposed Approach - Text Normalization

1. Normalization Process for hindi
   - 1.1 Spelling correction
   - 1.2 Ambiguous words
   - 1.3 Sounds
   - 1.4 Phontic words
   - 1.5 Transliteration
2. Dictionary method for marathi using bilingual lexicon

Proposed Approach - NLP and Sentiment Analysis

1. POS Tagging to identify nouns, adjectives and adverbs
   - 1.1 Use machine learning to figure out POS for hindi
   - 1.2 In case of marathi, use bilingual dictionary and then tag the statement
2. Adjective and Adverb extraction
   - 2.1 Easier to do with POS tagged statements.
   - 2.2 If POS doesn't work, then use lexicon lookup
   - 2.3 Use senti wordnet or HSWN for looking up polarity count
   - 2.4 If lookup fails, use extended wordnet api for getting sense
3. Negation and Classification using Classifier
   - 3.1 Figureout negation in statements and then invert the POS tagged adjectives and adverbs
   - 3.2 Semi supervised naive bayes/svm classifier to classify the polarity - or -
   - 3.3 Simple summation of polarity will give a result to. Compare the two results

Proposed Approach: Algorithm for language identification

Arguments

w : word to identify language for
sentence : sentence to which word belongs

```
if Leh is None:
        Leh = Qeh(Lhi) for i in Lh
if Lem is None:
        Lem = Qem(Lmi) for i in Lm
if not model:
        Model = CRF(FNGram(Leh), FNGram(Lem), FN-
Gram(Le))
        if not w in D:
                w = stem word(w)
                        if w not in D:
                                w = find most similar word(D,
w)
l = arg max(Model, sentence, w)
return l
```

Algorithm for pos tagged with negetation

Algorithm steps

if max(tags in sentence) is english:
TaggedSentence = POSTaggerEnglish(sentence)
else max(tags in sentence) is hindi:
TaggedSentence = POSTaggerHindi(sentence)
return replace negetive phrases with antonyms(TaggedSentence)

Algorithm for polarity identification

Algorithm steps

```
languageTaggedSentence = (LanguageIdentifier(sentence, word) for word in sentence).join(' ')
posTaggedSentence = POSTag(languageTaggedSentence)
polarWords = extractAdjectivesAdverbs(posTaggedSentence)
wordPolarity = dict()
for word in polarWords:
        if word tagged as english:
                wordPolarity[word] = sentiwordnetPolarity(word)
        elif word tagged as hindi:
                wordPolarity[word] = hindiSentiwordnetPolarity(word)
        else:
                hindiWword = hindiMarathiBilingualDictionary(word)
                wordPolarity[hindiWord] = hindiSentiwordnetPolarity(hindiWord)
return wordPolarity
```

Algorithm to classify polarity

Algorithm steps
Arguments
sentence : sentence to which word belongs
Variables and methods
NaiveBayesClassifier -¿ Trained to return polarity of the entire sentence given tokens with polarity values LinearCalculation -¿ Simple Summation based polarity classifier
wordDictionary = PolarityIndentification(sentence) return NaiveBayesClassifier(wordDictionary), LinearCalculation(wordDictionary)

Outline

Application

1. Reviews
2. Discourse Analysis
3. Feedback Analysis
4. Other areas

Outline

Conclusion

1. Accuracy of 95 percent on target
2. Imporvement of transliteration sentiment analysis in general
3. Usage of pre trained models over on the fly dynamic models for peformance gains

Outline

References


Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. "Hindi subjective lexicon: A lexical resource for hindi polarity classification". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*. 2012. URL: http://web2py.iiit.ac.in/research_centres/publications/download/inproceedings.pdf.a92646aa66336f21.4c5245432731322d3637332e706466.pdf (visited on 12/07/2015).


Bijoy Kundu and Swarup Chandra. "Automatic detection of English words in Benglish text: A statistical approach". In: *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on*. IEEE, 2012, pp. 1–4. ISBN: 1-4673-4367-6.


Namita Mittal et al. "Sentiment Analysis of Hindi Review based on Negation and Discourse Relation". In: *Sixth International Joint Conference on Natural Language Processing*. 2013, p. 45. URL: http://www.basantagarwal.com/wp-

content/uploads/2013/02/IJCNLP-WORKSHOP_W13-4306.pdf (visited on 12/07/2015).

Pooja Pandey and Sharvari Govilkar. "A Framework for Sentiment Analysis in Hindi using HSWN". In: *International Journal of Computer Applications* 119.19 (2015). URL: http://search.proquest.com/openview/1549cf02e0848335d2dd7a268e05f025/1?pq-origsite=gscholar (visited on 11/11/2015).

Shashank Sharma, PYKL Srinivas, and Rakesh Chandra Balabantaray. "Text normalization of code mix and sentiment analysis". In: *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*. IEEE, 2015, pp. 1468–1473. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7275819 (visited on 11/13/2015).