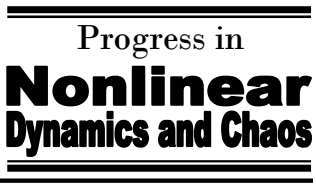
Progress in Nonlinear Dynamics and Chaos
Vol. 6, No. 2, 2018, 65-69
ISSN: 2321 – 9238 *(online)*
Published on 25 November 2018
www.researchmathsci.org
DOI: http://dx.doi.org/10.22457/pindac.v6n2a1



Review on Sentiment Lexicons of Indian Languages

Sajeetha Thavareesan

Department of Mathematics, Faculty of Science
Eastern University, Sri Lanka
Email: sajeethas@esn.ac.lk

Received 28 October 2018; accepted 24 November 2018

Abstract. Sentiment Analysis is the field of study which classifies the sentiments or emotions expressed in the human written text into positive and negative. In order to identify sentiments expressed in a text, sentiment lexical plays a crucial part. Sentiment analysis starts with prior polarity lexicons where entries are tagged with their prior scores. Most of the Sentiment analysis research works deal with English Sentiment lexicons but rare for Indian Languages like Tamil, Hindi, Urdu, etc. Some Sentiment analysis researches in Indian languages applied sentiment lexicons in English followed by translation from native language to English. In this approach authors missed language specific sentiment words and thus motivated to implement lexicons for Indian languages. This paper presents the sentiment lexicons presented in Indian languages.

Keywords: Sentiment Analysis; Sentiment Lexicon; Indian languages.

1. Introduction

Social media shifted objective applications to a subjective environment like opinion polls, surveys, etc. This expeditiously increase the digitized data of Indian Languages on the web. From NEWS to movie reviews sites, usage of Indian Languages like Tamil, Hindi, Bengali, etc. on the web is more than it ever was. Sentiment Analysis is the core of subjective environment which analyse movie reviews, feedbacks, etc.

Sentiment Analysis, one of the hot demanding research over the past few decades. Sentiment Analysis is the process of computationally identifying the opinions expressed in a piece of text to determine whether it is positive, negative, or neutral. Lexical analysis plays a decisive role in order to identify the sentiment. For example, words like love, hate, bad and beautiful directly indicate sentiment. So the development of good sentiment lexicons is necessary and compulsory while analysing sentiments based on lexicons.
SentiWordNet[1] for English is one of the important lexical resource which contains subjective polarity for each entity. But nowadays people prefer to give feedback in their native language. Analysing them demands language specific subjective lexicons. This served as the motivation for developing SentiWordNet for Indian Languages. SentiWordNet(s) are important resources for sentiment/opinion or emotion analysis task or Geospatial Information retrieval, Personalized search, Recommender System, Sentiment Tracking etc.

A lot of research attempts could be found in the literature for creation of Sentiment Lexicon in several languages and domains. In this paper, I discuss the generic

approaches followed for the development of SentiWordNet for Indian Languages using currently available resources in English.

2. Literature review

In the literature I found only six papers in sentiment lexicons in Indian languages. But the existing available systems are still do not meet the satisfaction level of end users'. The current trend is to attach prior polarity to the sentiment lexicon. Prior polarity is an approximation value collected from corpus and not exact which is a positive or negative score to a word out of context.

Das and Bandyopadhyay [2] proposed a template based online interactive game, called Dr Sentiment to create the PsychoSentiWordNet. The PsychoSentiWordNet is an extension of SentiWordNet 3.0 (Baccianella et. al. [3]), which holds human psychological knowledge along with sentiment knowledge. They considered Gender, Age, City, Country, Language and Profession as regulating aspects of human psychology. It supported 56 languages and therefore it was named as Global PsychoSentiWordNet. They asked the player to provide personal information. Dr Sentiment fetched random words from SentiWordNetsynsets and asked every player about his/her sentiment polarity towards those words. The gaming interface had four types of question templates. They performed concept- culture wise, Age wise, Gender wise, location- age wise, location- profession wise and gender- location wise analysis and found some interesting results. They observed that some words tagged as positive from one part and negative by other part of the world depending on culture. From their results sentiments were varied depending on age and also they observed that females were more positive than males.

Das and Bandyopadhyay [4] developed the prior polarity lexicons for three languages: Bengali, Telugu and Hindi. They developed based on two lexicons in English: SentiWordNet [11] and Subjectivity WordList [5]. They selected 8,427 words from SentiWordNet and neglected 2652 from Subjectivity WordList. They used the threshold of 0.4 to remove ambiguous entries and to avoid words got lost their subjectivity after translation. Also they removed the duplicates and the words tagged as "anypos" in Subjectivity WordList. They identified the stem words of inflected words and checked for the existence in the lexicon list. If it was found in the list, they added to the new list, otherwise just ignored. After that they calculated the prior polarity scores automatically with the help of the online program Dr. Sentiment as in [1]. From their study they found that, prior polarity values got influenced by geographical location, age, sex, profession, etc. NEWS and Blog corpus of Bengali was compared with Multi-Perspective Question Answering (MPQA) [5] and Internet Movie Database (IMDB) [6] datasets in English. They obtained 74.6% and 80.4% for Blog corpus whereas 72.16% and 76.0% for NEWS corpus as precision and recall respectively. The results indicate that the coverage of the Bengali SentiWordNet is reasonably good. They obtained 56.59% and 75.57% as precision for positive and negative. For Hindi 88% and 91% accuracy for positive and negative were obtained, and 82% and 78% for Telugu.

In [7] authors used four types of lexicons: SentiWordNet [11], Subjectivity Lexicon [5], AFINN111 lexicon and Opinion lexicon in English to develop SentiWordNet for Tamil. Lexicons about 16791 words with thresh hold above 0.4 were selected from SentiWordNet. Weakly subjective lexicons, lexicons tagged as "anypos" and words scored as neutral were also omitted from Subjectivity Lexicon. Among them they

selected the words which were tagged with same tag in both lists and removed the words tagged with conflicting tags. Altogether they had 15,823 words in new list. They added other two lexicons and for a given word presented more than once in the lexicon list, majority opinion was considered.

The solution for a word presented in two list with two opposing tag was not mentioned in their approach. Finally, Google translator was used to translate the new lexicon list to target language, Tamil. The problems they faced during translation were: Some words were not translated into Tamil because of lacks of such words, translation of multi word entries was a challenge- in a few cases a multi-word entry would get translated to an accurate single word and in some cases a single word entry would get translated to a multi-word entry. Final list of9495 words wereselected after validation by5 Tamil annotators. After validation 190 words were marked as ambiguous and 540 words as wrongly translated. Fleiss Kappa score for the final set was 0.663.

In [8], authors developed SentiWordNet(s) for three Indian languages: Bengali, Hindi and Telugu. They developed using multiple computational techniques like, WordNet based, dictionary based or corpus based.SentiWordNet and Subjectivity Word List were merged and the duplicates were removed to create the new sentiment lexicon then selected those whose orientation strength was above the heuristically identified threshold value of 0.4 and weakly subjective words were discarded from the Subjectivity word list. In the next stage the words with POS category in the Subjectivity word list was undefined and tagged as “anypos” were checked with the SentiWordNet list for validation. Inflected forms of a word in subjectivity list was clustered around the stem of the word and then checked in the SentiWordNet for validation. If the stem word was existed in the SentiWordNet then it was added to the new list. Otherwise it was discarded.

In Bilingual Dictionary Based Approach, English to Indian languages synsets developed under Project English to Indian Languages Machine Translation Systems (EILMT) was used to transfer each English synset/word in the merged sentiment lexicon into target language. Two English- Hindi electronic dictionaries: SHABD- KOSH[1] and Shabdanjali[2] had been identified for Hindi. These two dictionaries have been merged automatically by replacing the duplicates. The positive and negative sentiment scores for the Hindi words were copied from the English SentiWordNet. 22,708 Hindi word entries were in the final list. The bilingual dictionary based translation process had resulted in 35,805 Bengali entries and they were manually checked and removed 1688 words. Final list consisted of 34,117 words. Three English-Telugu dictionaries were merged by removing the duplicates and the positive and negative sentiment scores for the Telugu words were copied from English SentiWordNet. The dictionary based translation process had resulted in 30,889 Telugu entries. A WordNet based lexicon expansion strategy was adopted to increase the coverage of the new SentiWordNet(s). Hindi and Bengali WordNet(s) were publicly available but not for Telugu. Also they developed automatic antonym generation using sixteen hand crafted rules and corpus based approach was used to capture the language/culture specific words.A Conditional Random Field based model was trained with the seed list corpus along with multiple linguistics features such as morpheme, parts- of- speech, and chunk label which had been extracted by the shallow parsers. An intuitive game was developed to enhance the SentiWordNet(s).

In addition to this a Lexical Resource for Hindi Polarity Classification was developed by Bakliwal et al. [9]. Hindi subjectivity lexicons obtained 79.03% accuracy

on product reviews. In [10] Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets was done for Hindi and Marathi languages. Hindi and Marathi obtained 83.06% and 97.87% respectively for MultiDict.

3.Discussion

SentiWordNets for Indian Languages had been developed using Dr Sentiment, Bilingual dictionary, corpus and WordNet based approaches. PsychoSentiWordNet was developed for 56 languages but no evaluation was carried out as there were no corpora found. Except [9] and [10], others did not evaluated their lexicons. All the lexicons mentioned in this paper were categorized into positive and negative. Also authors used the same polarity values mentioned in the English SentiWordNet to the target language.

Table 1: Summary

No.	Reference	Lexicon	Source(s) used and accuracy obtained
1	[1]	56 languages	Dr. Sentiment (Online game)
2	[4]	Bengali, Hindi and Telugu	English SentiWordNet and Subjectivity Word List
3	[7]	Tamil	English SentiWordNet, Subjectivity Word List, AFINN111 lexicon and Opinion lexicon.
4	[8]	Bengali, Hindi and Telugu	English SentiWordNet, Subjectivity Word List, WordNets, Bilingual Dictionary and corpus
5	[9]	Hindi	WordNet, accuracy- 79.03%
6	[10]	Hindi and Marathi	Linked WordNets, Accuracy- 83.06% and 97.87% respectively.

4. Conclusion

Sentiment Analysis plays a crucial role in all sectors and thus researchers need lexicons which might incorporate slang words, transliteration words from other languages to native language, short forms of words, etc. Lexicon based Sentiment Analysis purely depends on the quality of the lexicons. There were no publicly available lexicons for some Indian languages. In [4] authors mentioned about publicly available Tamil WordNet but that is not accessible now. All the methods described here were based on validation by annotators followed by translation from English SentiWordNet. Also they used the same polarity values most of the time mentioned in English SentiWordNet. Some language or culture specific words could be missed out during the translation from English SentiWordNet. Thus, all the lexicons mentioned in this review paper need improvements in the prior polarity values. Also they need to incorporate all language or culture specific words into SentiWordNet and need to categorize them into positive, negative and neutral to increase the coverage of the sentiments.

I proposed a model to incorporate the above mentioned lacks in the existing lexicons to develop a Tamil sentiment lexicon. It is planned to incorporate slang words, short forms of words like

க க க போ (கருத்துக்களை கட்சிதமாக கவ்விக்கொள்கிறாய் போ) and transliteration words. Polarity values are planned to update with the help of ten linguistic experts.

REFERENCES

1. Baccianella, Stefano, A.Esuli and F.Sebastiani, Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, *Lrec*., 10 (2010) 2010.
2. A.Das and S.Bandyopadhyay, Dr Sentiment Knows Everything!, *Proceedings of the ACL-HLT 2011 System Demonstrations*, Portland, Oregon, USA, 21 June 2011 50–55.
3. B.Stefano, A.Esuli, and F.Sebastiani, SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *In the Proc. of LREC-10.*
4. A.Das and S.Bandyopadhyay, Dr Sentiment Creates SentiWordNet(s) for Indian Languages Involving Internet Population, *In the theIndoWordNet Workshop (ICON*), December, Kharagpur, India, 2010.
5. Mpqa.cs.pitt.edu. (2018). *Subjectivity Lexicon | MPQA*. [online] Available at: https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/ [Accessed 9 Nov. 2018].
6. IMDb. (2018). [online] Available at: https://www.imdb.com/interfaces/ [Accessed 9 Nov. 2018].
7. A.Kannan, G.Mohanty and R.Mamidi, Towards Building a SentiWordNet for Tamil, *Published December 2016 in ICON Proceedings of the 13th International Conference on Natural Language Processing.*
8. A.Das and S.Bandyopadhyay, SentiWordNet for Indian Languages, *Proceedings of the 8th Workshop on Asian Language Resources*, Beijing, China, 21-22 August 2010, pages 56–63.
9. A.Bakliwal, P.Arora and V.Varma, Hindi subjective lexicon: A lexical resource for Hindi polarity classification, *Int. J. Comput. Linguist. Appl.*, 2012.
10. A. R.Balamurali, A.Joshi, and P.Bhattacharyya, Cross-lingual sentiment analysis for Indian languages using linked wordnets, In COLING 2012.
11. SentiWordNet, (2018). [online] Available at: https://sentiwordnet.isti.cnr.it/ [Accessed 28 Nov. 2018].