International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

SENTIMENT ANALYSIS USING TELUGU SENTIWORDNET

DR. S. SURESH KUMAR, Assistant Professor & Head, Department of Information Technology,
JNTUH Universtiy College of Engineering, Jagtial, Telangana-505501.
Email: sureshsanampudi@gmail.com
PINNINTI HARITHA REDDY, M.TECH Student, Department of Information Technology,
JNTUH Universtiy College of Engineering, Jagtial, Telangana-505501.
Email:pinnintiharitha@gmail.com

ABSTRACT: In recent times, sentiment analysis in low resourced languages and regional languages has become emerging areas in natural language processing. Researchers have shown greater interest towards analyzing sentiment in Indian languages such as Hindi, Telugu, Tamil, Bengali, Malayalam, etc. In best of our knowledge, microscopic work has been reported till date towards Indian languages due to lack of annotated data set. In this paper, we proposed a two-phase sentiment analysis for Telugu news sentences using Telugu SentiWordNet. Initially, it identifies subjectivity classification where sentences are classified as subjective or objective. Objective sentences are treated as neutral sentiment as they don't carry any sentiment value. Next, Sentiment Classification has been done where the subjective sentences are further classified into positive and negative sentences. With the existing Telugu SentiWordNet, our proposed system attains an accuracy of 74% and 81% for subjectivity and sentiment classification respectively.

Keywords *– Natural Language Processing, Sentiment Analysis, Telugu, SentiWordNet, News sentences*
.

1. INTRODUCTION

In natural language processing (NLP), sentiment analysis is a technique that deals with analyzing the emotions, sentiments, opinions of an individual towards a product, movies, events, news or organizations, etc. [1]. The primary task of sentiment analysis is to identify the polarity of a text in a given document. The polarity may be either positive, negative or neutral. Sentiment analysis can be applied to text in three categories namely, sentence level, document level, and aspect level. Sentence level analysis focuses on identifying sentence-wise polarity value in a given document. Document level analysis determines the polarity value based on consideration of the whole document. In aspect level analysis, it identifies the polarity of every aspect (word-wise) in a given text. Telugu is the second most popular language in India after Hindi. According to Ethnologue list of most-spoken languages worldwide, Telugu ranks fifteenth in the list, and a total of 85 million Telugu native speakers exist across the world [2]. In the Telugu language, several e-Newspapers are available which publish news on a daily basis such as Eenadu, Sakshi, Andhrajyothy, Vaartha, and Andhrabhoomi, etc. SentiWordNet is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications [3].

According to Esuli and Sebastiani [3], "SentiWordNet is the result of the automatic annotation of all the synsets of WordNet towards the notions of positivity, negativity, and neutrality". Each synset is associated with three numerical scores pos(s), neg(s), and obj(s) which indicate "positive", "negative", and "objective" i.e., neutral respectively.

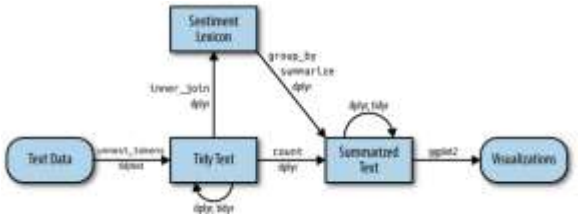

Fig.1: Example figure

In the recent past, researchers have shown their interest towards sentiment analysis in the context of Indian languages such as Hindi, Bengali, Telugu, Punjabi, Marathi, etc. [9]. Das and Bandyopadhyay [9] deployed a computational technique on English sentiment lexicons and English-Bengali bilingual dictionary to developed a Bengali SentiWordNet. In their subsequent work [10], they have extended their work and added two more Indian languages such as Hindi and Telugu to the SentiWordNet through an interactive gaming strategy called "Dr. Sentiment" to create and validate the SentiWordNet(s) for three

Indian languages with the help of Internet users. In this game, they considered SentiMentality analysis based on concept-culture wise, age wise and gender wise. Further, they have used this SentiWordNet to predict the polarity of a word and also suggested four approaches namely, the dictionary based, WordNet-based, corpus-based and interactive game (Dr. Sentiment) [11] to increase the coverage of generated SentiWordNet. In dictionary-based approach, they have developed a bilingual dictionary for English and Indian languages. In the Wordnet-based approach, they expanded the WordNet using synonym and antonym relations. In an automatic corpus-based approach, it captures the language/culture specific words to develop the corpus of SentWords. Finally, an interactive game is designed to identify the polarity of a word based on four questions which have to be answered by the users.

2. LITERATURE REVIEW

SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining:

In this work we present SE NTI WORD NET 3.0, a lexical resource explicitly devised for supporting sentiment classification and opinionmining applications. SENTIWORDNE T 3.0 is an improved version of SEN TIWORDNET 1.0, a lexical resource publicly available forresearch purposes, now currently licensed to more than 300 research groups and used in a variety of research projects worldwide. BothSEN TIWORDNET 1.0 and 3.0 are the result of automatically annotating all WORDNET synsets according to their degrees of positivity,negativity, and neutrality. SE NTIWORDNET 1.0 and 3.0 differ (a) in the versions of WOR DNET which they annotate (WORDNE T2.0 and 3.0, respectively), (b) in the algorithm used for automatically annotating WORDNE T, which now includes (additionally to theprevious semi-supervised learning step) a random-walk step for refining the scores. We here discuss SENTIWORD NET 3.0, especiallyfocussing on the improvements concerning aspect (b) that it embodies with respect to version 1.0. We also report the results of evaluatingSEN TIWORDNET 3.0 against a fragment of WOR DNET 3.0 manually annotated for positivity, negativity, and neutrality; these resultsindicate accuracy improvements of about 20% with respect to SEN TIWO RDNE T 1.0.

Lexicon-based methods for sentiment analysis:

We present a lexicon-based approach to extracting sentiment from text. The Semantic Orientation CALculator (SO-CAL) uses dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporates intensification and negation. SO-CAL is applied to the polarity classification task, the process of assigning a positive or negative label to a text that captures the text's opinion towards its main subject matter. We show that SO-CAL's performance is consistent across domains and in completely unseen data. Additionally, we describe the process of dictionary creation, and our use of Mechanical Turk to check dictionaries for consistency and reliability.

Dr sentiment creates SentiWordNet (s) for Indian languages involving internet population:

The discipline where sentiment / opinion / emotion has been identified and classified in human written text is well known as sentiment analysis. A typical computational approach to sentiment analysis starts with prior polarity lexicons where entries are tagged with their prior out of context polarity as human beings perceive using their cognitive knowledge. Till date, all research efforts found in sentiment analysis literature deal mostly with English texts. In this article, we propose an interactive gaming (Dr Sentiment) technology to create and validate SentiWordNet for three Indian languages, Bengali, Hindi and Telugu by involving Internet population. Dr Sentiment is an online game introduces a fictitious character, interact with players using series of questions and finally reveal the behavioral or sentimental status of any player and store the lexicons as the players polarized during playing. A number of automatic, semiautomatic and manual validations and evaluation methodologies have been adopted to measure the coverage and credibility of the developed SentiWordNet(s).

SentiWordNet for Indian languages:

The discipline where sentiment/ opinion/ emotion has been identified and classified in human written text is well known as sentiment analysis. A typical computational approach to sentiment analysis starts with prior polarity lexicons where entries are tagged with their prior out of context polarity as human beings perceive using their cognitive knowledge. Till

date, all research efforts found in sentiment lexicon literature deal mostly with English texts. In this article, we propose multiple computational techniques like, WordNet based, dictionary based, corpus based or generative approaches for generating SentiWordNet(s) for Indian languages. Currently, SentiWordNet(s) are being developed for three Indian languages: Bengali, Hindi and Telugu. An online intuitive game has been developed to create and validate the developed SentiWordNet(s) by involving Internet population. A number of automatic, semi-automatic and manual validations and evaluation methodologies have been adopted to measure the coverage and credibility of the developed SentiWordNet(s).

Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality:

Current sentiment analysis systems rely on static (context independent) sentiment lexica with proximity based fixed-point prior polarities. However, sentimentorientation changes with context and these lexical resources give no indication of which value to pick at what context. The general trend is to pick the highest one, but which that is may vary at context. To overcome the problems of the present proximity-based static sentiment lexicon techniques, the paper proposes a new way to represent sentiment knowledge in a Vector Space Model. This model can store dynamic prior polarity with varying contextual information. The representation of the sentiment knowledge in the Conceptual Spaces of distributional Semantics is termed Sentimantics.

3. METHODOLOGY

In recent times, sentiment analysis in low resourced languages and regional languages has become emerging areas in natural language processing. Researchers have shown greater interest towards analyzing sentiment in Indian languages such as Hindi, Telugu, Tamil, Bengali, Malayalam, etc. In best of our knowledge, microscopic work has been reported till date towards Indian languages due to lack of annotated data set.

In this paper using SentiWordNet author is detecting positive or negative sentences from Telugu sentences, this detection consists of two parts in which using first part we can detect objective or subjective from sentences and if objective words appear in the neutral list of SentiWordNet then that sentence will be consider as Neutral, if words not appear in SentiWordNet Neutral list then sentence words will check inside positive and negative list of SentiWordNet, if sentence words found in positive list then sentence will be consider as positive otherwise negative

Advantages of proposed system:

1. if sentences contains words from both positive and negative list then we take ratio of both positive and negative words and if positive ratio higher then sentence will be consider as positive else negative.

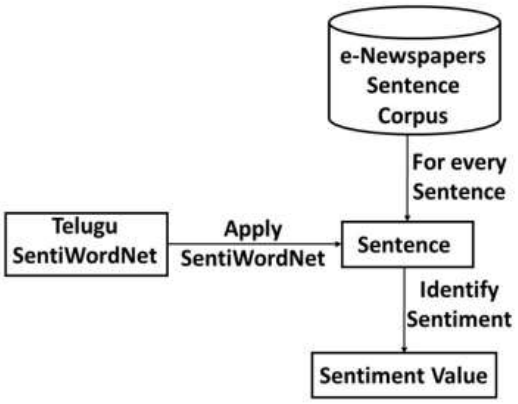

Fig.2: System architecture

MODULES:

In this project we have designed following modules

Data Collection & Annotation:

In this paper, data has been collected from the Telugu e-Newspapers namely, Eenadu, Sakshi, Andhrajyothy, Vaartha, and Andhrabhoomi, which are high rated newspapers in the states such as Andhra Pradesh and Telangana where the native language is Telugu. Our news dataset contains 1400 Telugu sentences from all the e-Newspapers as mentioned earlier ranging from the 1st of December 2016 to 31th of December 2016.

SentiWordNet for Sentiment Analysis:

SentiWordNet is a sentiment lexicon that associates the sentiment information to each and every word synset. We can represent SentiWordNet as Wordnet + sentiment information. In this paper, we have used Telugu Senti- WordNet [12-14] to perform the sentiment analysis. This SentiWordNet consists of

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

four files which contain negative, positive, neutral and ambiguous words respectively. The words in each file are categorized into five parts-of-speech tags namely, adjective

4. EXPERIMENTAL RESULTS

Fig.3: Home screen

Fig.4: Upload telugu sentiwordnet

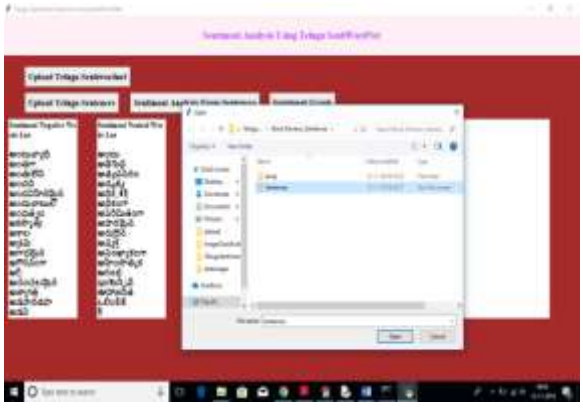

Fig.5: Upload telugu sentences

Fig.6: Sentiment analysis from sentences

Fig.7: Sentiment graph

5. CONCLUSION

The In Telugu languages, it's hard to find annotated dataset to perform NLP tasks such as POS tagging, sentiment analysis, sarcasm analysis, text summarization, etc. There are few annotated datasets available in this language. This paper exploits the available Telugu SentiWordNet to perform sentiment analysis for Telugu e-Newspapers sentences. The proposed system for sentiment analysis has attained an accuracy of 74% for subjectivity classification and 81% for sentiment classification in the domain of news data.

7. FUTURE SCOPE

In future, we need to improve the existing SentiWordNet to attains better accuracy and find an alternate way to make this SentiWordNet dynamic. It learns annotated data automatically and adds to the existing SentiWordNet.

REFERENCES

[1] Liu and Bing, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, 2012, pp. 1-167.

[2] Ethnologue Languages of the world [online]. Available: https://www.ethnologue.com/statistics/size
[3] Baccianella, Stefano, Andrea Esuli and Fabrizio Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," LREC, 2010, Vol. 10.
[4] Turney and Peter D, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002.
[5] Pang Bo, Lillian Lee and Shivakumar Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL 2nd conference on Empirical methods in natural language processing Association for Computational Linguistics, 2002, Vol. 10
[6] Pang Bo and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004.
[7] Hatzivassiloglou, Vasileios and Kathleen R. McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1997.
[8] Taboada and Maite, "Lexicon-based methods for sentiment analysis," Computational linguistics, 2011, pp. 267-307.
[9] Das, Amitava and Sivaji Bandyopadhyay, "Sentiwordnet for bangla," Knowledge Sharing Event-4: Task 2, 2010.
[10] Das, Amitava and S. Bandyopadhay, "Dr sentiment creates SentiWordNet (s) for Indian languages involving internet population," in Proceedings of Indo-wordnet workshop, 2010.
[11] Das, Amitava and Sivaji Bandyopadhyay, "SentiWordNet for Indian languages," in Asian Federation for Natural Language Processing, China, 2010, pp. 56-63.
[12] Das Amitava and Sivaji Bandyopadhyay, "Dr Sentiment knows everything!" in Proceedings of the 49th annual meeting of the association for computational linguistics, human language technologies, systems demonstrations, Association for Computational Linguistics, 2011.
[13] Das Amitava and Bjrn Gambck, "Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality," in Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, 2012.
[14] D Das, S Poria, CM Dasari and S Bandyopadhyay, "Building resources for multilingual affect analysis A case study on Hindi, Bengali and Telugu," Workshop Programme, 2012.
[15] BG Patra, D Das, A Das and R Prasath "Shared task on sentiment analysis in Indian languages (SAIL) tweets-an overview," in International Conference on Mining Intelligence and Knowledge Exploration, Springer International Publishing, 2015, vol. 9468.
[16] Kumar S.S., Premjith B., Kumar M.A. and Soman K.P, "AMRITA CEN-NLP@ SAIL2015 Sentiment analysis in Indian Language using regularized least square approach with randomized feature learning," in International Conference on Mining Intelligence and Knowledge Exploration, Springer International Publishing, 2015, vol. 9468.
[17] SS Prasad, J Kumar, DK Prabhakar and S Pal, "Sentiment Classification: An Approach for Indian Language Tweets Using Decision Tree," in International Conference on Mining Intelligence and Knowledge Exploration, Springer International Publishing, 2015, vol. 9468.
[18] Sarkar, Kamal and Saikat Chakraborty, "A sentiment analysis system for Indian language tweets," in International Conference on Mining Intelligence and Knowledge Exploration, Springer international Publishing, 2015, vol. 9468.
[19] Venugopalan Manju and Deepa Gupta, "Sentiment Classification for Hindi Tweets in a Constrained Environment Augmented Using Tweet Specific Features," in International Conference on Mining Intelligence and Knowledge Exploration, Springer International Publishing, 2015, vol. 9468.
[20] SS Mukku, N Choudhary and R Mamidi, "Enhanced Sentiment Classification of Telugu Text using ML Techniques," in 25th International Joint Conference on Artificial Intelligence, 2016.