Tweets and Telugu Senti WordNet Based on Sentimental Analysis using NLP

Dr.G.Shobana [1] | **D.Jhansi** [2] | **A. Siva Naga Manikanta Sai** [2] | **J. Vara Prasad** [2] | **G. Devi Prasanna Kumar Reddy** [2]

[1]Professor, Department of CSE, NRI Institute of Technology, India
[2]B.Tech Student, Department of CSE, NRI Institute of Technology, India

ABSTRACT

Recent years have witnessed a growth in the use of NLP for sentiment analysis in regional and low-resource languages. There has been a rise in researchers' dedication to the study of emotion analysis. No microlevel studies of human emotions have been published to far, perhaps owing to a lack of annotated data. Emotional recognition is an important study field that may provide many valuable insights. Expressions of feeling may be communicated in a number of outwardly obvious ways, including words, body language, and even typed text. The content-based classification issue at the heart of emotion identification in text documents makes use of ideas from natural language processing (NLP) and deep learning. A person's emotional state seems to be hard to determine from their written communication. Recognizing the tone of written communication is essential for effective human-computer interactions (HCI). The ability to express one's sentiments via one's words may be shown in a variety of ways, including through one's face, through text, and through one's voice. Since sufficient research has been done on face and spoken emotion recognition, a text-based emotional identification system also needs to attract academics. The capacity of computational linguistics to identify human emotions in text is useful in many practical contexts. In this project, we introduced a new approach to extracting specific user data from their tweets and other social media activity. For the most part centering on the subject's emotions, it will categorize the user's actions as sad, happy, or neutral. Several apps make advantage of the data we gather to deduce a person's emotional state.

KEY WORDS: Natural Language Processing, Sentiment Analysis, Telugu, SentiWordNet, News sentences

1. INTRODUCTION

Sentiment analysis is a method used in natural language processing (NLP) for deciphering an individual's feelings and thoughts on a certain topic, whether it a product, a piece of media, an event, an institution, etc. [1]. Sentiment analysis is used to determine the overall tone of a material. It's possible for the polarity to be positive, negative, or even neutral.

At the sentence level, the document level, and the aspect level, sentiment analysis may be applied to text. A document's polarity value may be determined by sentence using a sentence-level analysis. The polarity value is established after an in-depth study of the whole material has been performed. For use in aspect level analysis, it can tell you whether words in a text are positive or negative.

Telugu, after Hindi, is India's most widely spoken language. The number of native Telugu speakers is estimated to be 85 million [2]. On the Ethnologue list of most-spoken languages globally, Telugu comes in at number fifteen. There are several daily Telugu e-Newspapers, such as Eenadu, Sakshi, Andhrajyothy, Vaartha, Andhrabhoomi, etc.

SentiWordNet is a dictionary built specifically to aid with sentiment analysis and opinion mining [3]. When all the synsets of WordNet are automatically annotated with positive, negative, and neutral connotations, you get "SentiWordNet," as stated by Esuli and Sebastiani [3]. pos(s), neg(s), and obj(s) are three number ratings assigned to each synset that represent the "positive," "negative," and "objective," or neutral, connotations, respectively.

While there are a number of emotion analyzers available for the English language ([4-8]), not nearly as much has been done for Indian languages ([9-25]). Primarily, this is due to the paucity of supporting materials written in Indian languages.

In this study, we present a Telugu news sentiment analyzer that operates on a sentence-by-sentence basis. Subjectivity analysis is the first of two stages in the sentiment analysis process. In subjectivity analysis, we sort the sentences in the supplied corpus into subjective and objective categories. Furthermore, we evaluate if a subject's phrase is positive or negative. Since there is no emotional weight attached to an objective statement, it is classified as a neutral sentence. Because of this, in the first stage, the system determines if a statement is subjective (positive, negative) or objective (neutral). The second step of the process involves the computer determining whether the subjective statements are favorable or negative.

Here is how the remainder of the paper is structured: In Chapter 2, we discuss similar research. In Section 3, we detail the suggested framework for such a study of sentiment. In Section 4, we go into the experimental findings. Concluding thoughts and suggestions for the future are presented in Section 5.

2. LITERATURE SURVEY

Researchers have recently showed an interest in doing sentiment analysis in the context of Indian languages such as Hindi, Bengali, Telugu, Punjabi, and Marathi, amongst others [9-25]. Das and Bandyopadhyay [9]

constructed a Bengali SentiWordNet by using a computational approach to English emotion lexicons and an English-Bengali bilingual dictionary. Through the use of an interactive gaming strategy known as "Dr. Sentiment," the researchers were able to develop and validate the SentiWordNet(s) for three Indian languages with the assistance of Internet users. This was accomplished in their subsequent work [10], in which they extended their previous work and added two additional Indian languages to the SentiWordNet. These languages include Hindi and Telugu. Within the context of this game, concept-culture wise, age wise, and gender wise sentimentality analysis was taken into consideration.

In addition, they have used this SentiWordNet to forecast the polarity of a word. In addition, they have proposed four methods, namely, the dictionary-based, WordNet-based, corpus-based, and interactive game (Dr. Sentiment) [11], to enhance the coverage of the created SentiWordNet. They used an approach based on dictionaries and created a multilingual dictionary that covers both English and many Indian languages. Within the framework of the Wordnet-based methodology, they expanded the WordNet by including synonym and antonym associations. When using an automated technique that is based on a corpus, it records the linguistic and culturally specific terms in order to build a corpus of SentWords. Finally, an interactive game has been developed to determine the polarity of a word by basing its analysis on the players' responses to four questions about the word's meaning.

An alternative strategy for developing resources for multilingual affect analysis was given by Dipankar et al. [14] in the context of Indian languages. They utilized English as the source language while developing WordNet affixes for the three Indian languages of Hindi, Bengali, and Telugu. These affixes were created using English. They used WordNet, which is a database of every language and is accessible to the general public through the internet, in order to translate content into target languages.

SAIL stands for "Sentiment Analysis in Indian Languages," and Patra et al. [15] organized a shared task under that name with the intention of encouraging additional scholars to focus on sentiment analysis in Indian languages. During that meeting, a number of scholars presented their approach to analyzing

sentiment in Indian languages including Hindi, Bengali, Tamil, and others. [16-18]. The regularized least square technique with randomized feature learning was proposed by Kumar et al. [16] as a method for identifying sentiment inside the Twitter dataset. In a similar vein, Prasad et al. [17] suggested a sentiment analyzer for Hindi tweets that was based on a decision tree. The authors Sarkar et al. [18] built a sentiment analysis system for tweets written in Hindi and Bengali by using a multinomial naive Bayes classifier. The algorithm selects features based on the number of unigrams, bigrams, and trigrams in each tweet.

The sole study that has been published on Telugu sentiment analysis is done by Mukku et al. [20]. In order to train the Doc2Vec model, they made use of the raw corpus that was made available by the Indian Languages Corpora Initiative (ILCI). Additionally, they made use of the Doc2Vec tool that gives the semantic representation of a sentence in the dataset that was made available by Gensim, which is a Python module. Methods of machine learning, such as support vector machines, logistic regression, naive bayes, multi-layer perceptron neural networks, decision trees, and random forest classifiers, are used to train the system. They have carried out tests on binary and ternary classifications of feelings and opinions.

3.PROPOSED SYSTEM

We offered an innovative way for identifying an individual's personal information pertaining to tweets and social media activity inside the framework of the system that we designed. On the basis of this information, it will categorize the user's behavior as either sad, joyful, or in a neutral condition, concentrating mostly on the individual's sentiments. This information enables us to detect the emotions that a person is experiencing, and it is employed in a wide number of applications.
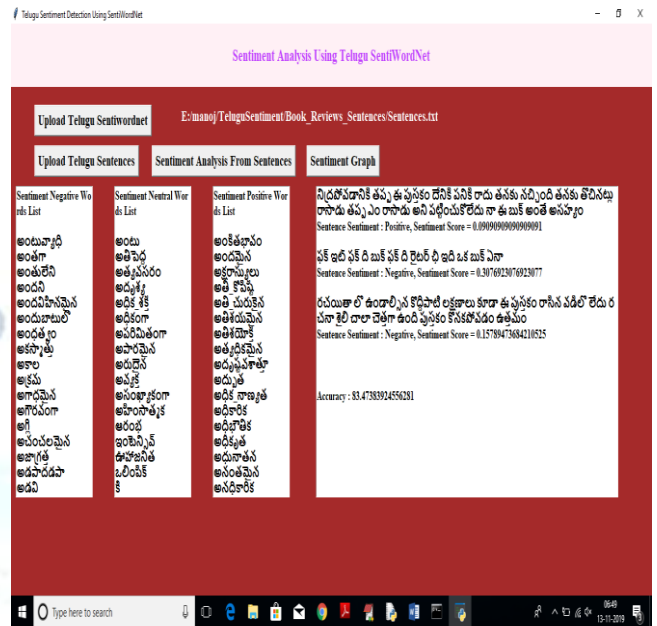
4. RESULTS

Figure 1: Comparison of accuracy with proposed and existing algorithm

Figure 2: Comparison of Accuracy with different models

5. CONCLUSION

It might be challenging to locate an annotated dataset in Telugu in order to carry out natural language processing (NLP) activities such as POS tagging, sentiment analysis, sarcasm analysis, text summarization, and so on. This language offers access to a limited number of datasets that have been annotated. This study makes use of the Telugu SentiWordNet that is currently available in order to do sentiment analysis on phrases taken from Telugu e-Newspapers. Within the realm of news data, the

suggested method for analyzing sentiment has achieved an accuracy of 74% for subjectivity classification and 81% for sentiment classification. These scores pertain to the study of sentiment rather than subjectivity. In the future, we will need to modify the current SentiWordNet so that it achieves a higher level of accuracy and locate an alternative approach to make this SentiWordNet dynamic. It does this by autonomously learning annotated data, which it then adds to the current SentiWordNet.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

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] B. Srinivasa Rao D. Vijaya Kumar and K. Kiran Kumar, "Power quality improvement using Cuckoo search based multilevel facts controller", Journal of Engg. Research Vol.10 No. (4A) pp. 252-261, DOI: 10.36909/jer.10895.

[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.