Sentiment Analysis using Telugu SentiWordNet

Reddy Naidu
Computer Science & Engineering
National Institute of Technology
Rourkela, Odisha - 769008
Email: naidureddy47@gmail.com

Santosh Kumar Bharti
Computer Science & Engineering
National Institute of Technology
Rourkela, Odisha - 769008
Email: sbharti1984@gmail.com

Korra Sathya Babu
Computer Science & Engineering
National Institute of Technology
Rourkela, Odisha - 769008
Email: prof.ksb@gmail.com

Ramesh Kumar Mohapatra
Computer Science & Engineering
National Institute of Technology
Rourkela, Odisha - 769008
Email: mohapatrark@nitrkl.ac.in

***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.**

***Index Terms*—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.

There exist several sentiment analyzers for the English language [4-8] but, in the context of Indian languages, little work has been done [9-25]. The primary reason behind is the lack of the available resources in Indian languages.

In this paper, we proposed a sentence-level sentiment analyzer for Telugu news. It is a two-step sentiment analysis process namely, subjectivity analysis and sentiment analysis. In subjectivity analysis, we classify the subjective and objective sentences from the given corpus. Further, we analyze the sentiment of subjective sentences either negative or positive. The objective sentences are treated as neutral sentences as it doesn't carry any sentiment value for the sentence. Therefore, in the first phase, the system classify the sentences as either subjective (positive, negative) or objective (neutral). In the second phase, the system classify the subjective sentences as either positive or negative.

The rest of the paper is organized as follows: Section 2 describes related work. Section 3 explains the proposed model for sentiment analysis. Experimental results are discussed in Section 4. Section 5 draws the conclusion with future work.

2. Related Work

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. [9-25]. 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.

In the context of Indian languages, Dipankar *et al.* [14] proposed an alternate way to build the resources for multilingual affect analysis. They have prepared WordNet affects for the three Indian languages such as Hindi, Bengali, and Telugu, and used English as a source language. For translation into target languages, they used WordNet of every language which is publicly available over the internet.

To motivate more researchers towards the sentiment analysis in Indian languages, Patra *et al.* [15] conducted a shared task called SAIL (Sentiment Analysis in Indian Languages). In that event, many researchers have presented their method to analyze sentiment in Indian language such as Hindi, Bengali, Tamil, etc. [16-18]. Kumar *et al.* [16] has suggested regularized least square approach with randomized feature learning to identify sentiment in the Twitter dataset. Similarly, Prasad *et al.* [17] proposed decision tree based sentiment analyzer for Hindi tweets. Sarkar *et al.* [18] developed a sentiment analysis system for Hindi and Bengali tweets using multinomial naive Bayes classifier that use unigrams, bigrams and trigrams for the selection of features.

Mukku *et.al.* [20] is the only reported work for Telugu sentiment analysis. They have used raw corpus provided by Indian Languages Corpora Initiative (ILCI) to train the Doc2Vec model and for pre-processing, Doc2Vec tool that gives the semantic representation of a sentence in the dataset provided by Gensim, a Python module. Machine learning techniques are used to train the system such as support vector machine, logistic regression, naive bayes, multi-layer perceptron neural network, decision tree and random forest classifiers. They have conducted experiments on binary and ternary sentiment classification.

3. Proposed Scheme

In this section, we proposed an automatic sentiment analyzer for Telugu e-Newspapers sentences. A model is shown in Figure 1. It starts with data collection and annotation. Further, using Telugu SentiWordNet, it classifies the sentiment of each sentence in news corpus. Finally, it compares the classification result with the manually annotated result for error analysis.

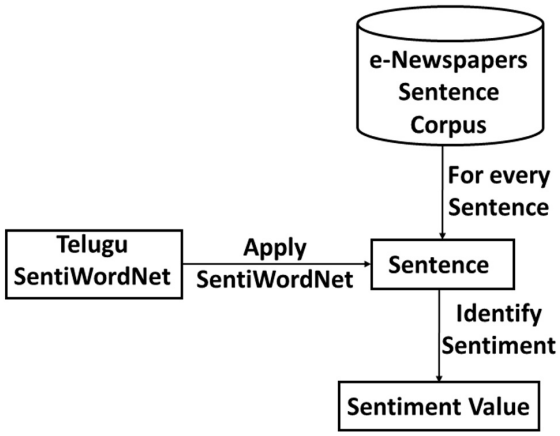

Figure 1: Model for sentiment analysis

3.1. 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 1^{st} of December 2016 to 31^{th} of December 2016. The number of sentences collected from each newspaper is shown in Table 3.

TABLE 1: List of e-Newspapers used for the data collection

	$Negative$	$Positive$	$Neutral$	$Total$
Eenadu	201	79	90	370
Sakshi	190	60	80	330
Andhrajyothy	137	55	58	250
Vaartha	144	50	56	250
Andhrabhoomi	100	52	48	200

This dataset is provided to the four annotators who have proficiency in the Telugu language, and belong to states of Andhra Pradesh and Telangana to annotate the sentiment of sentences in the dataset. They have interpreted the news sentences into three classes such as positive, negative, and

neutral. We approached the inter-annotator agreement using Cohen's kappa coefficient and got the annotation consistency (k value) to be 0.91. This manually annotated data is used as the baseline for comparison with system result.

3.2. 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 SentiWordNet [12-14] to perform the sentiment analysis. This SentiWordNet consists of 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 (a), noun (n), adverb (r), verb (v) and unknown (u). We have used neutral words file for the subjectivity classification, negative and positive words file for the sentiment classification. The list of words in the Telugu SentiWordNet and their categorization is shown in Table 3.

TABLE 2: Telugu SentiWordNet data categorization

	$Negative$	$Positive$	$Neutral$	$Ambiguous$
Adjective	1116	659	86	515
Noun	1066	544	124	320
Verb	833	363	60	156
Adverb	102	90	11	6
Unknown	959	480	78	96

3.2.1. Subjectivity Classification. Algorithm 1 explains the subjectivity classification which takes the corpus of Telugu news sentences as the input and outputs the subjective news sentences (SNS) file. It has performed by comparing each word in the sentence with the SentiWordNet neutral keywords file (neukf). If the word is present, the sentences are treated as objective sentences and discards in this level as they don't carry any sentiment value (neutral) and the remaining are treated as subjective sentences and stores in SNS file.

3.2.2. Sentiment Classification. Algorithm 2 explains the sentiment classification which takes the corpus of subjective news sentences (SNS) as the input and outputs the sentiment of a sentence. It has performed by comparing each word in the sentence with the SentiWordNet positive keywords file (poskf) and negative keywords file (negkf). If the word is present in poskf, the sentiment of that sentence is considered as positive, and if the word is present in negkf, the sentiment of that sentence is considered as negative. Otherwise, the sentence is simply discarded as any word of that sentence is not matched with any of the keywords in negkf and poskf.

In Algorithm 2, there is a high chance that some words in the sentence are matched with the negative keywords file, and some words in the same sentence are matched with positive keywords. In that scenario, it is hard to decide the sentiment of the sentence. To resolve this issue, we are keeping count variable to identify this kind of sentences. If the count is greater than one, the sentence is matched in both the lists $poskf$ and $negkf$. So, we are adopting sentiment score to identify the actual sentiment of a sentence. To find the sentiment score of the sentence, calculate the number of positive words (PWS) and negative words (NWS) in the same sentence. Then, calculate the positive ratio and negative ratio and Total sentiment score of the sentence using the equations 1, 2 and 3 respectively.

ALGORITHM 1: $Subjectivity\_Classification$

Input: Corpus of Telugu news headlines (C), SentiWordNet neutral keywords file ($neukf$)
Output: List of Subjective Sentences file (SNS)
Notation: C: corpus, S: sentence, TF: tokens file, T: token
Initialization : SNS = $\{\emptyset\}$
while S *in* C **do**
  $TF = get\_Tokens\ (S)$
  for T *in* TF **do**
    if *(* T *is present in* $neukf$ *)* **then**
      Sentence S is Objective (Neutral),
      Discard the sentence
    end
    else
      Sentiment is treated as Subjective
      Sentence $SNS \leftarrow SNS \cup S$
    end
  end
end

$$PR = \frac{PWS}{TWS} \tag{1}$$

$$NR = \frac{NWS}{TWS} \tag{2}$$

$$Sentiment\_Score = PR - NR \tag{3}$$

where,
PR= Positive Ratio, NR= Negative Ratio,
PWS= Number of Positive words in a given sentence,
NWS= Number of Negative words in a given sentence,
TWS= Number of words in a given sentence.

4. Experimental Results & Analysis

This section deals with the results obtained from the SentiWordNet approach. To experiment this, we have collected data from Telugu e-Newspapers and used Telugu SentiWordNet. The testing set consists of the 1400 sentences out of which 1068 are subjective, and the remaining 332 are objective sentences.

Initially, subjective classification was performed. It has correctly identified the 772 sentences (Tp) as subjective where the ground truth is 1068 and correctly identified the 275 sentences (Tn) as objective where the ground truth is

ALGORITHM 2: *Sentiment_Classification*

Input: Corpus of Telugu subjective news sentences (SNS),
SentiWordNet negative keywords file ($negkf$),
SentiWordNet positive keywords file ($poskf$)
Output: Sentiment of a news Sentence
Notation: SNS: corpus, S: sentence, TF: tokens file, T: token

```
while S in SNS do
    TF = get_Tokens (S)
    count = 0
    for T in TF do
        if ( T is present in poskf ) then
            Sentiment of S is Positive
            count = count + 1
        end
        else if ( T is present in negkf ) then
            Sentiment of S is Negative
            count = count + 1
        end
        else
            Sentence is treated as objective sentence
        end
    end
    if ( Count > 1) then
        Sent_S = Sentiment_Score(S)
        if (Sent_S > 0.0) then
            Sentiment of S is Positive
        end
        else
            Sentiment of S is Negative
        end
    end
end
```

332. The Fp is 57, which are objective but classified as subjective and Fn is 296, which are subjective but classified as objective.

In the next step, sentiment classification was performed. The 772 subjective sentences are considered out of which 262 are positive and 510 are negative. It has correctly identified the 202 sentences as positive (Tp), where the ground truth is 262 and correctly identified the 427 sentences as negative (Tn), where the ground truth is 510. The Fn is 60, which are negative but classified as positive and Fp is 83, which are positive but classified as negative. All these parameters are shown in Table 3.

TABLE 3: Results in terms of Confusion Matrix

	Tp	Fn	Fp	Tn
Subjectivity Classification	772	296	57	275
Sentiment Classification	202	60	83	427

There are three statistical parameters namely, $precision$, $recall$ and $F-score$ are also evaluated to test the performance of the experimented work using the equations 4, 5 and 6 respectively. The results are shown in terms of statistical parameters for subjectivity classification and sentiment classification in Table 4.

$$Precision = \frac{T_p}{T_p + F_p} \tag{4}$$

$$Recall = \frac{T_p}{T_p + F_n} \tag{5}$$

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

where,
T_p = true positive, F_p = false positive, F_n = false negative.

TABLE 4: Results in terms of $Accuracy$, $Precision$, $Recall$, $F-score$

	$Accuracy$	$Precision$	$Recall$	$F-score$
Subj_Class	74%	0.93	0.722	0.812
Senti_Class	81%	0.708	0.770	0.737

where,
Subj_Class = Subjectivity Classification,
Senti_Class = Sentiment Classification.

To obtain the confusion matrix as shown in Table 3, we used human annotated sentiment values as ground truth. The ground truth values are as follows:

- Total sentences in test data set : 1400
- Subjective sentences: 1068
- Total positive sentences: 653 and negative sentences: 415

Based on the above ground truth, error analysis is shown in Table 3 through Confusion matrix. This result entirely depends on the quality of SentiWordNet. The obtained accuracy can be improved by improving the Telugu SentiWordNet. In this work, we haven't used any machine learning techniques to analyze the performance since there is no direct provision to apply on SentiWordNet.

5. Conclusion & Future Work

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.

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.

Acknowledgments

The authors would like to thank Bala Prakash, Manikanta, Vijay and Madhusudan for annotating the collected dataset. All the annotators are native to the states of Andhra Pradesh & Telangana and have a good knowledge of the Telugu language.

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 I*nternational 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.