IJARIIT

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

A rule based sentiment analysis in Telugu

Pravarsha Jonnalagadda
jpravarsha.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Krishna Pratheek Hari
hkrishnapratheek.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Sandeep Batha
bathasandeep.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Hashresh Boyina
bhashresh.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

ABSTRACT

Sentimental Analysis in the English language is a relatively easier task to perform as it has a predefined set of rules followed and accepted universally. But, when it comes to Indian languages, there isn't a benchmark dataset. Moreover, if a dataset exists, it cannot be validated as a similar sentence may differ in the meaning as the regional languages are very unpredictable and have no proper rules. In this paper, we used a Rule-Based Approach for Telugu sentiment analysis using Telugu SentiWordNet. Here, we obtained the sentiment using SentiWordNet and validated the results using ACTSA which is an annotated corpus data set.

Keywords— *SentiWordNet, SentiPhraseNet, PoS tagger*

1. INTRODUCTION

Natural language processing is an area of computer science and artificial intelligence which deals with the interactions between human and computer languages. The sentimental analysis is an important part of Natural Language Processing and it is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [1]. It helps us in understanding the sentiments, in most cases the opinions.

The sentimental analysis is important as its applications have spread to almost every possible domain, from consumer products, services, healthcare, and financial services to social events and political elections. It also can be applied to text in three categories namely aspect level, sentence level and document level. In aspect level analysis, the polarity of every aspect (word-wise) in a given text is obtained. Sentence level analysis helps in identifying sentence-wise polarity value in a given document. Document-level analysis determines the polarity value based on consideration of the whole document.

Telugu, despite being a regional language in India is ranked 15th in the list of most widely spoken languages with over 81 million native speakers worldwide [2]. Hence there is an absolute necessity to analyse the sentiments of this language.

2. RELATED WORK

Researchers have shown their interest in sentiment analysis in the context of Indian languages such as Hindi, Malayalam, Telugu, Odia, Marathi, etc. In Malayalam, the corpus is collected from the Malayalam websites to do the sentiment classification. But the major problem with the corpus is spelling errors in user's feedback which will immensely affect the accuracy of the analysis. A rule-based approach is proposed by Deepu S. Nair and Co. [3] for finding the sentiment of Malayalam text from the film review websites *i.e.*, from the users' feedback whether the sentiments obtained is either positive, negative or neutral from their writings.

In Odia, Sahu *et al*. [4] suggested an empirical study of supervised learning techniques to classify Odia movie reviews. Trying to analyze the sentiments of Odia people expressed in Odia movie reviews, a system that classifies the Odia text in positive and negative sentiment using supervised classification techniques has been developed. Python language was used to write the program. They have considered three supervised classifiers namely, Naive Bayes, Support Vector Machine and Logistic Regression and followed the NLTK framework to perform the task.

To motivate more researchers towards the sentiment analysis in Indian languages, Patra *et al* [5] conducted a shared task called SAIL (Sentiment Analysis in Indian Languages). In that event, many researchers have presented their method to analyse sentiment in Indian language such as Hindi, Bengali, Tamil etc. Kumar *et al* [6] have suggested regularized least square approach with randomized feature learning to identify sentiment in the Twitter dataset. Similarly, Prasad *et al* [7] proposed decision tree based sentiment analyser for Hindi tweets. Sarkar *et al* [8] developed a sentiment analysis system for Hindi and Bengali tweets using multinomial naïve Bayes classifier that uses unigrams, bigrams and trigrams for the selection of features.

For Telugu language, Naidu *et al* [9] proposed a two-phase sentiment analysis for Telugu news sentences using Telugu SentiWordNet. Initially, the subjectivity classification was done where sentences are classified as subjective or objective. Objective sentences are treated as neutral sentiment as they don't carry any sentimental value. Next, Sentiment Classification has been done where the subjective sentences are further classified into positive and negative sentences.

3. PROPOSED WORK

In this section, we will see the proposed work and process flow to do the sentiment analysis in the Telugu language. The System Model has been depicted in figure 1.

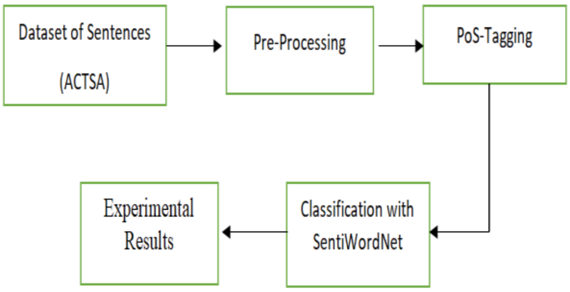

Fig. 1: A rule-based approach

3.1 Data-Collection

In this paper, we have followed the ACTSA (Annotated Corpus for Telugu Sentiment Analysis) which was developed by Mukku *et al*. [10]. They have and the dataset particulars have been shown in Table 1. Most of the corpora available for Sentiment Analysis is harvested from sources like review data from e-commerce websites where customers express their opinion on products freely and posts from social networking sites like Twitter and Facebook.

3.2 Pre-Processing

In this step, PoS Tagger is used to split up the sentences into different parts of speech as well as to identify and remove stop words etc. The PoS tagger which is used is in this paper is "Sivareddy PoS tagger"[11], developed by Mr Sivareddy, which is specifically developed for Telugu language. The Tags used in this paper have been mentioned in Table 2.

Table 1: Statistics about the ACTSA dataset

News articles	321
Cleaned Sentences	11952
Objective Sentences (Removed)	4327
Uncertain Sentences (Removed)	1802
Disagreement Sentences	512
------- Classified	99
------- Removed	413
Positive Sentences	1489
Negative Sentences	1441
Neutral sentences	2475
Total sentences	5410

Table 2: Representation of parts of speech in PoS tagger

Parts of Speech	Associated Tag
Noun	NN
Verb	VM
Adverb	RB
Adjective	JJ

After the PoS Tagging, we have extracted Nouns, Adverbs, Adjectives and Verbs as they are the words which provide the sentiment in a given sentence. The output contains the following columns separated by tab space.

3.3 Sentiment classification using SentiWordNet

In this step, SentiWordNet is used to find the sentiment of words that are obtained from ACTSA. SentiWordNet can be understood as a dictionary which contains the sentiment of a word along with the meaning.

ALGORITHM 1: *Sentiment Classification using SentiWordNet*

```
Input : Nouns, Adjectives, Adverbs, nouns from sentences in
        ACTSA corpus
Output: Positive, Negative and Neutral keywords files
Notation: LOWF: List of words files, ifile: file in LOWF, ofile: file in
          SentiWordNet, iword: word in ifile, oword: word in ofile
for ifile in LOWF
 | for iword in ifile
 |  | for ofile in SentiWordNet
 |  |  | for oword in ofile
 |  |  |  | if iword==oword then
 |  |  |  |  | write iword to outputfile with same name as ifile
 |  |  |  |  | flag=1
 |  |  |  |  | break
 |  |  | for flag==1 then
 |  |  |    break
 |  | if flag==0 then
 |  |  | write iword to unknown words file
 |  | else
 |  |  | flag=0
```

3.4 Sentiment classification using TextBlob

Since the Telugu SentiWordNet is a finite dictionary of words, there are a lot of words in ACTSA whose sentiments are not specified in the SentiWordNet.

ALGORITHM 2: *Sentiment Classification using TextBlob*

```
Input : unknown words file from SentiWordNet classification
Output: Positive, Negative and Neutral keywords files
Notation: UWF: unknown words files, blob: object of
          TextBlob class, tran_blob: translated word,
          pol: polarity
for word in UWF
 | blob = TextBlob(word)
 | tran_blob = translate the blob to English
 | pol = polarity of blob
 | if pol > 0.0 then
 |  | write blob, tran_blob, and polarity to pos.txt
 | else if pol < 0.0 then
 |  | write blob, tran_blob and polarity to neg.txt
 | else
 |  | write blob, tran_blob and polarity to neu.txt
```

This results in a lot of unknown words, whose sentiment is to be identified. For this purpose, we use TextBlob, an online package, which contains some methods that help in finding the sentiment of the English equivalent of the particular Telugu word.

Despite using TextBlob, there are many other complicated words that are unable to get translated into English. Hence, the sentiment of these words was identified by finding their meaning and giving them proper polarity accordingly.

3.5 Finding Accuracy

Finally, we put together all the words and divided them into 3 files namely, positive, negative, and neutral files based on their polarity. Since objective words should not be considered we only considered only positive and negative words.

To find out the legitimacy of this extended SentiWordNet we validated it with the annotated dataset i.e., ACTSA.

With the help of the confusion matrix, we obtain the Accuracy, F-measure, Precision, and Recall using the formulas below,

$$\text{Accuracy} = ((T_p+T_n)/(T_p+T_n+F_p+F_n)) \quad (1)$$

$$\text{Precision} = T_p/(T_p+F_p) \quad (2)$$

$$\text{Recall} = T_p/(T_p+F_n) \quad (3)$$

$$\text{F-measure} = (2*\text{Precision}*\text{Recall})/(\text{Precision} + \text{Recall}) \quad (4)$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

Table 3: Confusion Matrix

T_p = 412	F_p = 185
F_n = 142	T_n = 556

Where T_p= True Positive, T_n= True Negative, F_p= False Positive, F_n= False Negative.

Table 4: Tabulated Results

Accuracy (%)	Precision	Recall	F-Measure
74.74	0.69	0.74	0.71

ALGORITHM 3: *Finding Confusion Matrix*

```
Input : positive file and negative file obtained from combining
        outputs of all sentiment classifications
Output: Values of Confusion Matrix
Notation: LOF: List of files, ifile: file in LOWF, iword:
          word in ifile, tp: True Positive, fp: False Positive,
          fn: False Negative, tn: True Negative, oline: line in ACTSA
for oline in ACTSA
 if oline starts with '+' or '-' then
  for file in LOF
   for iword in ifile
    if iword is in oline and iword is not in ambiguous then
     if "pos" is in filename and oline starts with '+' then
      if iword is in completed then
       if value of iword in completed is not "tp" then
          append iword to ambiguous
          tp--
      else
       add (iword:"tp") to completed
       tp++
     if "neg" is in filename and oline starts with '+' then
      if iword is in completed then
       if value of iword in completed is not "fn" then
          append iword to ambiguous
          fn--
      else
       add (iword:"fn") to completed
       fn++
     if "pos" is in filename and oline starts with '-' then
      if iword is in completed then
       if value of iword in completed is not "fp" then
          append iword to ambiguous
          fp--
      else
       add (iword:"fp") to completed
       fp++
     if "neg" is in filename and oline starts with '-' then
      if iword is in completed then
       if value of iword in completed is not "tn" then
          append iword to ambiguous
          tn--
      else
       add (iword:"tn") to completed
       tn++
```

The confusion matrix is obtained using algorithm 2 and is tabulated in Table 3. Four statistical parameters namely accuracy, precision, recall and F-score are evaluated to test the performance of the experimented work using the Eqns. (1), (2), (3) and (4) respectively. These are tabulated in Table 4.

5. FUTURE WORK AND CONCLUSION

The availability of an annotated dataset has reduced the difficulty of Natural Language Processing in the Telugu language to some extent.

As we proposed in our paper, the SentiWordNet can be extended for as long as it encounters new words which are not specified in the SentiWordNet. But, even then there are possibilities for the existence of some drawbacks as one word may not always give appropriate sentiment for the whole sentence.

Though the accuracy obtained is74.74%, the above-mentioned flaw in our approach can be reduced by the usage of Bi-grams and Tri-grams, which can be an extension to the current approach.

6. ACKNOWLEDGEMENTS

The authors would like to thank Mr Reddy Naidu for his constant support and guidance in the process of working on this paper.

7. REFERENCES

[1] Liu and Bing, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, 2012, pp. 1–167.

[2] Liu and Bing, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, 2012, pp. 1–167.

[3] Deepu S. Nair, Jisha P. Jayan, Rajeev R.R, Elizabeth Sherly, "SentiMa - Sentiment Extraction for Malayalam", IIITM, Kerala

[4] Sanjib Kumar Sahu, Priyanka Behera, D. P. Mohapatra, Rakesh Chandra Balabantaray, "Sentiment analysis for Odia language using supervised classifier: information retrieval in Indian language initiative", SPECIAL ISSUE REDSET 2016 OF CSIT.

[5] Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath, "Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview", Springer International Publishing Switzerland 2015.

[6] S. S. Kumar, B. Premjith, M. A. Kumar, and K. P. Soman, "AMRITA_CEN-NLP@ SAIL2015 Sentiment analysis in Indian Language using regularized least squares approach with randomized feature learning," in International Conference on Mining Intelligence and Knowledge Exploration, Springer International Publishing, 2015, vol. 9468.

[7] S. S. Prasad, J. Kumar, D. K. 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.

[8] Kamal Sarkar 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.

[9] Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu and Ramesh Kumar Mohapatra "Sentimental Analysis using SentiWordNet" in IEEE WiSPNET 2017 conference.

[10] Sandeep Sricharan Mukku and Radhika Mamidi, "ACTSA: Annotated Corpus for Telugu Sentiment Analysis", Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, pages 54–58, Copenhagen, Denmark, September 8, 2017.

[11] Sivareddy, "Cross-Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources", "Proceedings of the Fifth International Workshop On Cross Lingual Information Access", 2011, pages 11-19.

