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

SentiPhraseNet: An extended SentiWordNet approach for Telugu sentiment analysis

Abhinav Garapati
gabhinav.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Naveen Bora
bnaveen.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Hanisha Balla
hballa.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Mohan Sai
bmohansai.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 data set 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 to build SentiPhraseNet. Here, we obtained the sentiment using SentiPhraseNet and validated the results using ACTSA which is an annotated corpus data set.

Keywords— *NLP, Sentiment analysis, POS tagging, Rule-based, SentiWordNet, SentiPhraseNet*

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.

A 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. The 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.
The rest of the paper is organized as follows: Section II describes related work. Section III explains the proposed model for sentiment analysis. Experimental results are discussed in Section IV. Section V draws the conclusion with future work.

2. SENTIWORDNET

A SentiWordnet is a set of unigram words with the associated sentiment. It is a sentiment lexicon associated sentiment information to each wordnet synset.

SentiWordnet = Wordnet + Sentiment Information

For each wordnet synsets, the following information is available in SentiWordnet
Positive Score pos(s)
Negative Score neg(s)
Objective Score obj(s)

We observed a few drawbacks with SentiWordNet.

2.1 Drawbacks of SentiWordNet

It has a fixed number of words in the list that makes it unable to identify the sentiment value of the testing sentences with unknown words.

	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

To deal with morphology while testing sentences, a morphological analyser such as "stemmer" is required for finding the root word.

#	Telugu Sentence	English Meaning
1	రాము మామిడిపండు **తింటాడు**	Ramu eats mango
2	రాము **తినడానికి** సిద్ధంగా ఉన్నాడు	Ramu is ready to eat
3	నేను రోజు అరటిపండు **తింటాను**	Every day I eat banana
4	నేను నిన్న దోశ **తిన్నాను**	Yesterday I ate dosha
5	పవన్ రేపు చికెన్ **తినబోతున్నాడు**	Tomorrow pavan will eat chicken
6	రాము **తింటున్నాడు**	Ramu is eating

It consists of unigram words that are unable to identify the correct sentiment in various situations.

#	Telugu Sentences	SentiWordNet	SentiPhraseNet
1	సైబర్ నేరాల పై **అవగాహన తక్కువే**, డిజిపి సాంబశివరావు వెల్లడి	అవగాహన (Awareness) (Positive)	అవగాహన తక్కువే (Negative)
2	నగదు రహిత చెల్లింపులపై **అవగాహన కార్యక్రమాలు**, కళాశాలలకు, విశ్వవిద్యాలయాలకు యూజీసీ సూచన	అవగాహన (Awareness) (Positive)	అవగాహన కార్యక్రమాలు (Positive)
3	అది ప్రక్రియలను, ఫలితాలను **నిర్లక్ష్యం చేసింది**.	నిర్లక్ష్యం (Neglect) (Negative)	నిర్లక్ష్యం చేసింది (Negative)
4	నీరు ప్రగతిని **నిర్లక్ష్యం చెయ్యొద్దు**, సీఎం చంద్రబాబు వెల్లడి	నిర్లక్ష్యం (Negative)	నిర్లక్ష్యం చెయ్యొద్దు (Positive)
5	**దార్శనిక నాయకురాలని కోల్పోయాం**, జయలలిత మృతికి పారిశ్రామిక సంఘాల సంతాపం	దార్శనిక (Philosophical) (Positive)	దార్శనిక నాయకురాలని కోల్పోయాం (Negative)
6	**గిరిజనుల త్యాగాలకు న్యాయమెక్కడ**?చంద్రబాబు సర్కారుపై ధ్వజమెత్తిన జగన్	త్యాగాలకు (Sacrifice) (Positive)	గిరిజనుల త్యాగాలకు న్యాయమెక్కడ (Negative)
7	ఆదాయం 10 లక్షలు దాటితే **గ్యాస్ రాయితీ ఉండదు**, పన్ను చెల్లింపుదారుల వివరాలు పెట్రోశాఖకు ఇవ్వనున్న ఐటి విభాగం	రాయితీ (Positive)	గ్యాస్ రాయితీ ఉండదు (Negative)
8	**పెరిగిన చలి తీవ్రత**, వాతావరణ శాఖ వెల్లడి	పెరిగిన (Positive)	పెరిగిన చలి తీవ్రత (Negative)

It contains a list of ambiguous words which are unable to predict sentiment properly

#	Telugu Sentences	SentiWordNet	SentiPhraseNet
1	ఓటుకి నోటు కేసులో **ఆరోపణ ఎదుర్కొన్న చంద్రబాబు**	ఆరోపణ (Allegation) (Ambiguous)	ఆరోపణ ఎదుర్కొన్న చంద్రబాబు (Allegations faced by chandrababu) (Negative)
2	చంద్రబాబు మీదున్న **ఆరోపణ ఋజువు చేస్తే** దేనికైనా సిద్ధం అన్న తెలుగుదేశం పార్టీ	ఆరోపణ (Allegation) (Ambiguous)	ఆరోపణ ఋజువు చేస్తే (If proves the allegations) (Positive)
3	పచ్చ కళ్ళు, **పిచ్చి రాతలు**, కృష్ణపట్నం పై ఈనాడు కట్టుకథలు, సైసార్సీపీ మండిపాటు	పిచ్చి (Madness) (Ambiguous)	పిచ్చి రాతలు (Mad Writings) (Negative)
4	చంద్రబాబు అంటే నాకు **పిచ్చి ప్రేమ**, తెలిపిన మంత్రి ఉమా మహేశ్వరరావు	పిచ్చి (Madness) (Ambiguous)	పిచ్చి ప్రేమ (Madness of love) (Positive)

With these drawbacks, SentiWordNet must be replaced. So, we build a SentiPhraseNet which consists of a combination of words that is, bigrams and trigrams. We tried to eliminate a few drawbacks of SentiWordNet in SentiPhraseNet.

3. 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 sentiment value. Next, Sentiment Classification has been done where the subjective sentences are further classified into positive and negative sentences.

4. 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.

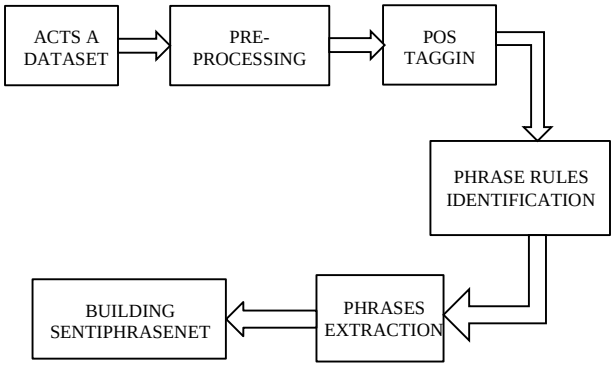

Fig. 1: A rule-based approach

4.1 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.

4.2 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 "Siva Reddy PoS tagger"[11], developed by Mr Sivareddy, which is specifically developed for the 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 ------- Classified ------- Removed	**512** **99** **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.

4.3 Identified phrase rules for SentiPhraseNet

After obtaining the POS tags for each word, we identified some phrase rules that is a combination of words, which gives sentiment. We found some rules for bigrams and trigrams.

Bigram Rules:

Noun + Adjective
Adjective + Noun
Adverb + Noun
Noun + Adverb
Adjective + verb
Adverb + Verb

Trigram Rules:

Adjective + Adjective + Noun
Verb + Adjective + Adjective
Noun + Noun + Verb
Noun + Verb + Noun
Verb + Noun + Noun

We only considered these rules because we observed that only these combinations contains higher sentiment that affects the polarity of the statement.

ALGORITHM 1: *Sentiment Classification using SentiPhraseNet*

```
Input : Nouns, Adjectives, Adverbs, nouns from sentences in
ACTSA corpus
Output: Positive, Negative and Neutral phrases files
Notation: LOWF: List of phrases files, ifile: file in LOWF, ofile: file in
SentiphraseNet, i: phrase in ifile, ophrase: phrase in ofile
for ifile in LOWF
    for iphrase in ifile
        for ofile in SentiPhraseNet
            for ophrase in ofile
                if iphrase==ophrase then
                    write iphrase to outputfile with same name as ifile
                    flag=1
                    break
            for flag==1 then
                break
        if flag==0 then
            write iphrase to unknown phrase file
        else
            flag=0
```

4.4 Sentiment Classification using TextBlob

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

This results in a lot of unknown phrases 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 phrase.

ALGORITHM 2: *Sentiment Classification using TextBlob*

```
Input : unknown phrases file from SentiPhraseNet classification
Output: Positive, Negative and Neutral phrases files
Notation: UWF: unknown phrases files, blob: object of
          TextBlob class, tran_blob: translated word,
pol: polarity
for word in UWF
    blob = TextBlob(phrases)
    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
```

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

4.5 Finding Accuracy

Finally, we put together all the phrases and divided them into 3 files namely, positive, negative, and neutral files based on their polarity.

Since objective phrases should not be considered we only considered positive and negative phrases. To find out the legitimacy of this extended SentiPhraseNet 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,

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

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

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

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

5. EXPERIMENTAL RESULTS AND ANALYSIS

Table 3: Confusion Matrix

$\mathbf{T_p}$ = 594	$\mathbf{F_p}$ = 200
$\mathbf{F_n}$ = 110	$\mathbf{T_n}$ = 441

Where,
$\mathbf{T_p}$= True Positive, $\mathbf{T_n}$= True Negative,
$\mathbf{F_p}$= False Positive, $\mathbf{F_n}$= False Negative.

Table 4: Tabulated Results

Accuracy (%)	Precision	Recall	F-Measure
77.002	0.74	0.84	0.79

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, iphrase:
          phrase 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 iphrase in ifile
        if iphrase is in oline and iphrase is not in ambiguous then
          if "pos" is in filename and oline starts with '+' then
            if iphrase is in completed then
              if value of iphrase in completed is not "tp" then
                append iphrase to ambiguous
                tp--
            else
              add (iphrase:"tp") to completed
              tp++
          if "neg" is in filename and oline starts with '+' then
            if iphrase is in completed then
              if value of iphrase in completed is not "fn" then
                append iphrase to ambiguous
                fn--
            else
              add (iphrase:"fn") to completed
              fn++
          if "pos" is in filename and oline starts with '-' then
            if iphrase is in completed then
              if value of iphrase in completed is not "fp" then
                append iphrase to ambiguous
                fp--
            else
              add (iphrase:"fp") to completed
              fp++
          if "neg" is in filename and oline starts with '-' then
            if iphrase is in completed then
              if value of iphrase in completed is not "tn" then
                append iphrase to ambiguous
                tn--
            else
              add (iphrase:"tn") to completed
              tn++
```

6. 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 SentiPhraseNet can be extended for as long as it encounters new phrases which are not specified in the SentiPhraseNet. As there are many phrases that are not in Sentiphrasenet, whenever a new phrase encounters then with the help of textblob we define its polarity and add it accordingly. By this way, we can achieve dynamism.

Though the accuracy obtained is 77.002%, the above-mentioned flaw in our approach can be reduced by the increasing number of rules and dynamism.

7. ACKNOWLEDGEMENTS

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

8. 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: An 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@ SAIL 2015 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.

