Building SentiPhraseNet for Sentiment Analysis in Telugu

Reddy Naidu
Department of CSE
ANITS, Sangivalasa
Visakhapatnam - 531162
naidureddy47@gmail.com

Santosh Kumar Bharti
Department of CSE
Pandit Deendayal Petroleum University
Gandhinagar, Gujarat - 382421
sbharti1984@gmail.com

Korra Sathya Babu
Department of CSE
National Institute of Technology
Rourkela - 769008
prof.ksb@gmail.com

***Abstract*—Sentiment analysis of Indian languages is a challenging task due to rich morphology and little availability of the annotated datasets. For languages like Hindi, Telugu, Tamil, Bengali, Malayalam, etc., SentiWordNets (SWNets) were developed to tag the sentiment of each word. In this article, we observed that some unigram words of the existing Telugu SWNet are classified as ambiguous and are not sufficient to analyze the sentiment. In such situations, bigram and trigram phrases can be used to resolve the problem of ambiguity in sentiment prediction. Therefore, we proposed an algorithm to build the Telugu SentiPhraseNet (SPNet) for the sentiment analysis in Telugu. To build SPNet, we have collected the data from various sources namely, Telugu e-Newspapers, Twitter and NLTK Indian Telugu data which resolves the problems with existing SWNet. With the proposed SPNet, we have performed the sentiment analysis and it is compared with SWNet, various existing Machine Learning approaches namely, Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), Multilayer Perceptron Neural Network (MLPNN), Decision Tree (DT) and Random Forest (RF). The performance of the proposed system outperformed the other existing approaches and attains an accuracy of 85.6%.**

Keywords*—*Natural Language Processing*, *SentiPhraseNet*, *Sentiment Analysis*, *Telugu*, *News Data*, *NLTK Indian Data

I. Introduction

Sentiment analysis is an application of NLP which deals with the identification of people's sentiments, emotions, and opinions towards a target such as products, services, events, movies, news, organizations, individuals, etc. [1]. It is a type of subjectivity analysis that focuses on identifying positive and negative opinions, emotions, and evaluations expressed in natural language [2]. This analysis helps the people to know what other people think or feel about their products, services, events, etc.

In recent times, sentiment analysis in Indian languages such as Hindi, Telugu, Bengali, Tamil, etc., has become an emerging area for research. It is due to the influence of native languages in communication. A massive amount of regional language data is getting generated every day. These data are in the form of tweets, reviews, news, blogs, movies, etc. According to Ethnologue list [3], Telugu ranks sixteenth of most-spoken languages worldwide. Among the regional languages in India, Telugu is a popular language and has approximately 75 million native speakers in India [4]. The regional data generated by Telugu communicating users stand just after Hindi among Indian languages. The absence of sufficient annotated Telugu resources and its morphological complexity pose sentiment analysis a challenging task.

#	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

Fig. 1: Different Morphological Structures of Word "eat" in Telugu Sentences

Recently, a Telugu SWNet [A bags-of-words which is labeled with corresponding sentiment value $i.e.$, either positive, negative or neutral.] was developed by Das *et al.* [5] to analyze the sentiment in Telugu data. This SWNet consists of a fixed number of unigram words and has the following limitations:

1) 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.
2) To deal with morphology while testing sentences, a morphological analyzer such as "stemmer" is required for finding the root word. An example of morphology in Telugu sentences is shown in Fig. 1.
3) It consists of unigram words that are unable to identify the correct sentiment in various situations as shown in Fig. 2.
4) It contains a list of ambiguous words which are unable to predict sentiment properly. An example is shown in Fig. 3.

An instance of morphological structure in the Telugu language is shown in Fig. 1. For the six Telugu sentences that appear in Fig. 1, the root word is "eat". The morphological existence of the root word is underlined.

It can be observed that only unigram words are not sufficient to predict the sentiment in a Telugu sentence. Sometimes, it requires additional support from surrounding words in the form of bigrams and trigrams to classify sentiment of Telugu sentences. Such situations are explained using various examples as shown in Fig. 2.

In Fig. 2, consider the first and second sentences. The Telugu meaning of 'awareness' is present in the positive list of Telugu SWNet. Therefore, the predicted sentiment in the first sentence is positive. However, the actual sentiment of the sentence is negative (when we consider a bigram of 'awareness' with the adjacent word as indicated in underlined bold). Hence, by using SWNet, the given sentence is misclas-

978-1-5386-8235-7/18/$31.00 ©2018 IEEE

#	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**, సిఎం 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)

Fig. 2: Various Situations where Bigram and Trigram Phrases are Preferable over Unigram Words for Sentiment Analysis

sified. Similarly, for the third and fourth sentences, the Telugu meaning of 'neglected' is listed in the negative list, but the real sentiment of the fourth sentence is positive, which is again misclassified by Telugu SWNet. For the sentences 5 to 8 of Fig. 2, misclassification is resolved using trigram phrases.

In case of ambiguous words in Telugu SWNet, the problem of sentiment analysis can be resolved using the context of the surrounding words. Such instances are shown in Fig. 3. It shows that the word 'allegation' can be used to classify a sentence as positive (first sentence) as well as negative (second sentence) using either bigram or trigram combinations. With the considerations of the above limitations of existing Telugu SWNet (excluding the morphological structure), we have proposed an algorithm to build a Telugu SPNet [A bags-of-phrases which are labeled with corresponding sentiment value i.e either positive, negative or neutral.]

The rest of the article is organized as follows: Section II describes related work. The proposed scheme is discussed in Section III and the Results and discussions are mentioned in Section IV. Finally, the conclusion & future work of the article is given in Section V.

#	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** దనికైనా సిద్ధం అన్న తెలుగుదేశం పార్టీ	ఆరోపణ (Allegation) (Ambiguous)	ఆరోపణ ఋజువు చేస్తే (If proves the allegations) (Positive)
3	పచ్చ కళ్ళు, **పిచ్చి రాతలు**, కృష్ణపట్నం పై ఈనాడు కట్టుకథలు, వైసార్సీపీ 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)

Fig. 3: Various Situations where Ambiguous Words from SWNet are replaced by Bigram and Trigram Phrases to resolve the Ambiguity

II. Related Work

This section gives a glimpse of the literature survey on sentiment analysis in low-resourced languages such as Hindi, Telugu, Tamil, Bengali, Arabic, Indonesian, etc. The major chunk of the work in the area of sentiment analysis is done in the English scripted language as it is the most dominating language across the world. In the domain of low-resourced languages, limited works have been reported so far. Due to the unavailability of annotated dataset for training, very few authors have worked in this direction [6] - [17].

A. Sentiment Analysis for Low Resourced Languages

Availability of annotated datasets for low resourced language is almost meagre. An unsupervised approach has been proposed by Wan *et al.* [6] for sentiment analysis in Chinese language using bilingual knowledge and ensemble techniques. In the absence of Chinese annotated data, the authors used machine translation features to translate the Chinese reviews into English and then identified the sentiment polarity value of the English reviews. Further, the ensemble methods are employed to combine the individual analysis results. An Indonesian tweet sentiment analyzer was developed by Aliandu *et al.* [7] using Naive Bayes approach. The authors had used unigram, Term-Frequency (TF), and Term Frequency-Inverse Document Frequency (TF-IDF) as the feature set to build the sentiment classification model. During feature extraction process, they haven't considered some tweet parameters namely, re-tweets, URL links, Twitter usernames, etc. To build the sentiment lexicons, Chen *et al.* [8] suggested a technique with the help of knowledge graph construction and comparison of graph propagation with label propagation for the languages such as Arabic, German, English, Italian, Japanese and Chinese.

B. Sentiment Analysis for Indian Languages

In the context of sentiment analysis in Indian languages, an event (shared task on Sentiment Analysis for Indian Languages (SAIL) was conducted by Patra *et al.* [9] on the developments in Indian language sentiment analysis. In this event, a good number of researchers participated with sentiment analysis in Indian languages. A decision tree (C4.5) based approach was suggested by Prasad *et al.* [10]. Similarly, Kumar *et al.* [11] proposed Regularized Least Square (RLS) approach with Randomized Feature Learning (RFL) for sentiment analysis in Hindi, Bengali and Tamil tweets. The tweets dataset was taken from SAIL 2015. A multinomial Naive Bayes classifier was suggested by Sarkar *et al.* [12] for sentiment analysis of Indian language tweets, and they considered unigrams, bigrams and trigrams as the feature set for training.

To build the SWNet for Indian languages, Das *et al.* [13] proposed a gaming strategy called "Dr Sentiment" with the involvement of internet population. They have considered age-wise, gender-wise and concept-culture wise analysis to identify the sentimental behavior of the people. Another concept was proposed by the same authors employing the four approaches namely dictionary based, WordNet-based, corpus-based, and gaming methodology [5]. In the dictionary-based approach, word-level translation process followed by error reduction technique has been adopted. A WordNet-based lexicon expansion strategy has been adopted to increase the coverage of the generated SWNet. To capture the language/culture specific words, a corpus-based approach has been adopted. The intuitive gaming methodology was adopted to creates a multilingual SWNets automatically.

Venugopalan *et al.* [14] has proposed a technique to capture the tweet specific features by calculating TF-IDF scores of unigrams for feature extraction to identify sentiment in Hindi tweets. Ravi *et al.* [15] has developed a Hinglish text [A blend of Hindi and English, in particular a variety of English used by speakers of Hindi, characterized by frequent use of Hindi vocabulary or constructions.] sentiment analyzer by employing Gain Ratio (GR) and Radial Basis Functional Neural Network (RBFNN) for feature selection. Nair *et al.* [16] has suggested Sentiment Extraction for Malayalam language (SentiMa) by using a rule-based approach with

the user's feedback of movies from the Malayalam movie review websites. Sentiment analysis in Odia language has been suggested by Sahu *et al.* [17] using NB classifier to perform the sentiment classification.

To the best of our knowledge, very little work has been reported on sentiment analysis so far on Telugu data. Mukku *et al.* [18] considered the Telugu news dataset to perform the sentiment analysis using raw corpus provided by Indian Languages Corpora Initiative (ILCI) which is used to train the Doc2Vec model. For pre-processing, they used Doc2Vec that gives the semantic representation of a sentence in the dataset provided by Gensim, a Python module. Various machine learning techniques are used to train and test the system such as SVM, LR, DT, NB, RF, and MLPNN. Similarly, Naidu *et al.* [19] has performed the sentiment analysis using existing Telugu SentiWordNet. They have proposed a two-phase sentiment analysis namely Subjectivity classification and sentiment classification for Telugu news sentences using Telugu SentiWordNet and attained an accuracy of 74% and 81% for subjectivity and sentiment classification respectively. Further, Mukku *et al.* [20] proposed a gold-standard annotated corpus named ACTSA (Annotated Corpus for Telugu Sentiment Analysis) has a collection of Telugu sentences taken from different Telugu news websites to support sentiment analysis in the Telugu Language.

III. Proposed Scheme

This section describes the proposed framework for building Telugu SPNet to identify sentiment in Telugu sentences as shown in Fig. 4. The framework starts with the process of data collection. The collected data is provided to the Telugu Parts-of-Speech (POS) tagger to assign suitable POS information for every token. Subsequently, an algorithm was deployed for phrase extraction from tagged data.

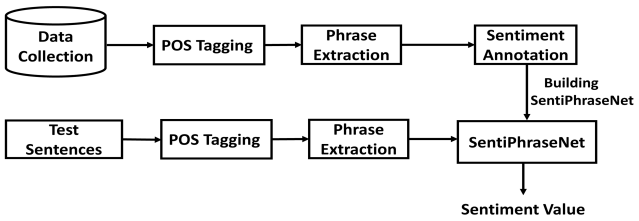

Fig. 4: Framework for Sentiment Identification in Telugu Data

Manual annotation was done to classify the sentiment of the extracted phrases for building the SPNet. Finally, the SPNet was used to predict the sentiment value of given sentences in the testing set.

A. Data Collection

In this research, we have collected data from various sources such as Telugu e-Newspapers, Twitter and NLTK Indian Telugu Data for building SPNet. The details of the datasets are given in Table I. It consists of 5000 Telugu sentences in the form of news headlines, tweets, and articles.

B. POS Tagging

POS tagging is the process that divides input sentences into atomic words and assigns them with corresponding POS information to each word based on their relationship with adjacent and related words in a sentence. In this paper, a Hidden Markov Model (HMM) based Telugu POS tagger [21] is used to identify the correct POS tag information. This Telugu tagger is monolingual tagger which uses target transition and emission probabilities estimated from the existing tagger. They followed Indian language standard (ILS) tagset [22] that comprise 21 tags to build the tagger. The tagger is trained using large data containing 3,152,199 tokens. It is capable of handling a large vocabulary, and also can predict the tags of unknown words using known words.

C. Phrase Extraction for Building SPNet

For building the SPNet, initially, we used a corpus of 5000 Telugu sentences for phrase extraction using Algorithm 1. It takes a Telugu sentence as an input from the corpus and identifies the POS tag value of each word. Further, it stores the tag value in the temporary tagged file (TF). Next, for the combination of every tag in TF, it checks the presence of bigram and trigram tags patterns such as (ADV + NN) || (ADJ + V) || (NN + ADJ) || (ADJ + NN) || (NN + ADV) || (ADV + V) || (ADV + ADJ + NN) || (V + ADV + ADJ) || (V + NN + NN) || (NN + V + NN) || (NN + NN + V). If any of the pattern (either bigram or trigram) is matched, it stores the matched pattern into a pattern set (PatSet) and extracts the phrase value of that PatSet into the sentence-wise set of phrases (SPS). This process is repeated for all the sentences in the corpus and stores all the corresponding extracted phrases into phrase file (PF). If none of the patterns matched, then the input sentence is classified as neutral.

D. Manual Annotation of Extracted Phrases

The extracted set of phrases is provided to the four human annotators who have good knowledge of the Telugu language and are native to the states of Andhra Pradesh and Telangana. They have annotated these phrases into three classes such as positive, negative, and neutral. In this research, Fleiss' Kappa coefficient is used to find the Inter-annotator Agreement (IAA) [23] as it is more suitable as the number of annotators are more than two. Fleiss' Kappa can be calculated using Equation 1. For this experiment, the attained IAA value is 0.89.

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

where,
$\bar{P} - \bar{P}_e$: gives the degree of agreement actually achieved above chance,
$1 - \bar{P}_e$: gives the degree of agreement that is attainable above chance.

TABLE I: Dataset used for Building the SPNet

Category of Dataset	Number of Sentences
News Headlines	2606
Eenadu	800
Sakshi	800
Andhrajyothy	600
Vaartha	406
Twitter Data	1400
NLTK Indian Telugu Data	994

The annotation result for sentiment classification of the phrases to build SPNet is given in Table II and a sample list of annotated phrase are shown in figures 5 and 6. The Proposed SentiPhraseNet is available to the researchers on Request.

TABLE II: Manual Sentiment Annotation of Generated Phrases

	Negative	Positive	Neutral
Bigram Phrases	480	628	1155
Trigram Phrases	1780	1613	1557

Algorithm 1: *Extraction_of_Phrases_for_SPNet(EPS)*

Input: Corpus of Telugu Sentences (C)
Output: List of bigram and trigram phrases
Notation: ADJ: Adjective,V: Verb, ADV: Adverb, NN: Noun, S: sentence, C: corpus, TF: tagged file, T: tag, SPS: Sentence-wise set of phrase, PF: Phrase file
Initialization : PF = {∅}
while S *in* C **do**
  $TF = Find\_POS\_Tag\ (S)$
  while T *in* TF **do**
    if *(Pattern of (ADV +NN) || (ADJ +V) || (NN +ADJ) || (ADJ +NN) || (NN + ADV) || (ADV + V) || (ADV + ADJ + NN) || (V + ADJ + NN) || (V + ADV + ADJ) || (V + NN + NN) || (NN + V + NN) || (NN + NN + V) is Matched)* **then**
      $\langle PatSet \rangle = Extract\_matched\_pattern\_phrase\ (TF)$
      $\langle SPS \rangle = Phrase\,[PatSet]$
    end
    else
      Input Sentence is Neutral
    end
  end
  $PF \leftarrow PF \cup \langle SPS \rangle$
end

Negative	Positive	Neutral
శోకసంద్రమైన తమిళనాడు (Tamilanadu is in Grief)	అభిమానంగా పలకరిస్తున్నారు (Greeting Affectionately)	వ్యక్తిగా మోడీ (Modi as Human Being)
కన్నీరుమున్నీరుగా విలపిస్తున్న (Weeping with distress)	సంతోషంగానే ఉంది (Feeling Happy)	తదుపరి ప్రధాన (The Next Main)
దేశవ్యాప్తంగా నిషేదించింది (Ban across the country)	అరుదైన రికార్డును (Rare Record)	మిగిలిన దేశాలకు (To the remaining countries)
తీవ్రంగా అవమానిస్తున్నారు (Belittling seriously)	మెరుగైన సౌకర్యాలు (Enhanced Facilities)	ఒక ప్రకటనలో (In one Advertisement)
తీవ్రంగా నష్టపోతుందని (Lose Badly)	ప్రత్యేక సౌకర్యాలు (Special Facilities)	వ్యక్తిగతంగా హాజరు (Attended in Person)
ఇబ్బందికరంగా మారుతున్నాయి (Becoming embarrassing)	ముఖ్యులతో ప్రత్యేకంగా (Special with Prime people)	ఒక గ్రామాన్ని (One Village)
విద్వేషపూరితమైన చర్యలు (Hateful Actions)	ఆసక్తికరమైన విషయాలు (Interesting Things)	ప్రతి విషయంలో (In every thing)

Fig. 5: Sample List of the Extracted Bigram Phrases of SentiPhraseNet

Negative	Positive	Neutral
జయలలిత ఓటమి పాలయ్యారని (Assuming Jayalalithaa has lost)	రాజకీయ తెరపై తిరుగులేని (Unstoppable in political life)	ప్రకటించిన టైమ్స్ సంపాదకులు (Declared by TIMES Editor)
జయలలిత మృతి తట్టుకోలేక (Unbearable feeling for losing Jayalalithaa)	తొందరపాటు నిర్ణయం కాదు (Not a hasty decision)	తెలిపిన మంత్రి గంటా (Specified by minister Gantaa)
ఆగ్రహం వ్యక్తం చేశారు (Expressed Anger)	ప్రదర్శనపై నిషేధం ఎత్తివేత (Lift ban on Performance)	మోడీని కలిసిన సెల్వం (Selvam met Modi)
సంభవించిన పెను భూకంపం (Occurred major earthquake)	ఆనందం వ్యక్తం చేసిన (Expressed happiness)	ఉర్జిత్ పటేల్ ప్రకటించారు (urjith Patel Announced)
అడుగంటుతున్న తుంగభద్ర నది (Scarcity in Tungabhadra river)	ప్రజలకు అవగాహన కల్పించేందుకు (To create public awareness)	ప్రధాని మోడీ పిర్కొన్నారు (Told by PM Modi)
నెలకొన్న గందరగోళ పరిస్థితులు (Formed confused situations)	నిబంధనలను సరళతరం చేసింది (Has simplified the rules)	కరుణానిధి ఇంటికి వెళ్లి (Went to karunanidhi's home)
భయాందోళనలు సృష్టించేందుకు కుట్రలు (Conspiracies to create fears)	ప్రథమ స్థానం సంపాదించుకున్నారని (Acquired the first place)	ఎయిరుపోర్ట్ అధికారులు తెలిపారు (Specified by Airport Officials)

Fig. 6: Sample List of the Extracted Trigram Phrases for SentiPhraseNet

E. Sentiment Classification using SPNet

To analyze the sentiment in Telugu sentences using SPNet, an algorithm is proposed and is shown in Algorithm 2. It takes testing set (C) and SPNet as the input to identify the sentiment class of the sentences in C. In the first step, the algorithm finds the sentence-wise POS tag information of sentences in C and stores it in tagged file (TF). Next, for the combination of every tag in TF, it checks the existence of the bigram and trigram tag patterns such as (ADV +NN) || (ADJ +V) || (NN +ADJ) || (ADJ +NN) || (NN + ADV) || (ADV + V) || (ADV + ADJ + NN) || (V + ADJ + NN) || (V + ADV + ADJ) || (V + NN + NN) || (NN + V + NN) || (NN + NN + V) and counts the number of total phrases (either bigrams or trigrams) in a sentence. All the matched pattern sets and its corresponding phrases are extracted in PatSet and SPS respectively.

Algorithm 2: *Sentiment_Classification_using_SPNet*

Input: Testing set of Telugu sentences (C1),$SPNet$
Output: $classification := positive, negativeorneutral$
Notation: ADJ: Adjective, V : Verb, ADV : Adverb, NN: Noun, S: sentence, T: Tag, TF: Tagged file, $PatSet$: Pattern set, SPS: Sentence-wise set of phrases, (C1): Testing set, $UKPL$: Unknown phrase list, SC: Sentiment score
Initialization : $UKPL$ = {∅}
while S *in* (C1) **do**
  $TF = Find\_POS\_Tag\ (S)$
  $countP = 0$
  while T *in* TF **do**
    if *(Pattern of (ADV +NN) || (ADJ +V) || (NN +ADJ) || (ADJ +NN) || (NN + ADV) || (ADV + V) || (ADV + ADJ + NN) || (V + ADV + ADJ) || (V + NN + NN) || (NN + V + NN) || (NN + NN + V) is Matched)* **then**
      $\langle PatSet \rangle = Extract\_matched\_pattern\_phrase\ (TF)$
      $\langle SPS \rangle = Phrase\,[PatSet]$
    end
  end
  $N_c = 0$, $P_c = 0$, $Neu_c = 0$
  while *(phrase in SPS)* **do**
    check the presence of the phrase in SPNet.
    if *(phrase present in positive list)* **then**
      $P_c \leftarrow P_c + 1$
    end
    else if *(phrase present in negative list)* **then**
      $N_c \leftarrow N_c + 1$
    end
    else if *(phrase present in neutral list)* **then**
      $Neu_c \leftarrow Neu_c + 1$
    end
  end
  if *(* $CountP == N_c$*)* **then**
    Given sentence is classified as negative.
  end
  else if *(* $CountP == P_c$ *)* **then**
    Given sentence is classified as positive.
  end
  else if *(* $CountP == Neu_c$*)* **then**
    Given sentence is classified as neutral
  end
  else if *((*$P_c > 0$*) && (*$Neu_c > 0$*) && (*$N_c == 0$*))* **then**
    Given sentence is classified as positive.
  end
  else if *((*$N_c > 0$*) && (*$Neu_c > 0$*) && (*$P_c == 0$*))* **then**
    Given sentence is classified as negative.
  end
  else if *((*$P_c > 0$*) && (*$N_c > 0$*) &&(*$Neu_c == 0$*))* **then**
    $SC = Find\_Sentiment_Score\ (SPS, PatSet)$
    if *(*$SC > 0$*)* **then**
      Given sentence is classified as positive
    end
    else if *(*$SC < 0$*)* **then**
      Given sentence is classified as negative
    end
    else
      Given sentence is classified as ambiguous.
    end
  end
  else
    $UKPL \leftarrow UKPL \cup SPS$
  end
end

For every phrase in SPS, it checks for its presence in SPNet (either negative, positive or neutral list). If all the phrases of a particular sentence exists in the positive list then a sentence is classified as positive, if all the phrases exist in the negative list then a sentence is classified as negative, and if all the phrases exist in the neutral list, then the sentence is classified as neutral. Similarly, if some of the phrases of a particular sentence is present in the negative list and some

are in neutral then the sentence is classified as negative. If some of the phrases are present in positive list, and some are in neutral then the sentence is classified as positive. If some of the phrases are present in negative list and some are in positive then calculate the sentiment score (SC) to classify the sentiment. The procedure to calculate SC is given in Algorithm 3. If SC is greater than zero, it is classified as positive, and if SC is less than zero, then a sentence is classified as negative. If SC is equal to zero, the sentence is classified as ambiguous.

Algorithm 3: $Sentiment\_Score\_Calculation$

```
Input: SPS, PatSet, SPNet
Output: Sentiment score
Notation: S: sentence, SPS: sentences-wise set of phrases, SC: Sentiment
  score, ToP: Tags of Phrase
Initialization : Dict = f(ADV : 2); (ADJ : 3); (NN : 1); (V : 1),
  W_p = {0} , W_n = {0}
while (phrase in SPS) do
    ToP = Extract_Tagset (PatSet)
    if phrase is present in positive list of either bigram or trigram then
        W = argmax_{j∈Dict}(ToP[j])
        if W > W_n then
            W_n = W
        end
    end
    else
        W1 = argmax_{j∈Dict}(ToP[j])
        if W1 > W_n then
            W_n = W
        end
    end
end
SC = W_p - W_n
```

Algorithm 3 takes SPS, PatSet, and SPNet as the input to calculate sentiment score of a particular sentence. It initializes a dictionary (Dict) with $(key, value)$ pair where key is the tag and value is the corresponding weight in the phrase. Here, four tags are considered with weight values namely adjective (ADJ) with weight = 3, adverb (ADV) with weight = 2, verb (V) with weight = 1 and noun (NN) with weight = 1. Next, for every phrase in (SPS), extract tags of phrase (ToP) using $(PatSet)$. Next, it checks the presence of that particular phrase in positive or negative lists of SPNet. For all the positive phrases in (SPS), calculate the maximum weight of the tags in the phrase using Dict and store in W_p. Similarly, for all the negative phrases in (SPS), calculate the maximum weight of the tags in phrase Dict and store in W_n. Finally calculates the SC by subtracting W_p and W_n.

IV. Results and Discussion

This section describes the performance of the proposed SPNet to identify the sentiment value for the Telugu news sentences. It starts with an experimental setup for deploying the proposed algorithm followed by evaluation parameters to check the performance of the proposed system. Further, it explains the experimental results and its discussion.

For testing the performance of proposed system, we have taken 8000 Telugu sentences in three categories namely, tweets, news articles and historical articles. The performance of the proposed SPNet has attained an accuracy of 85.6%. The accuracy can be increased further by applying the dynamic nature to extend the SentiPhraseNet. The performance of sentiment analysis is compared with existing machine learning techniques [18] as well as SWNet [19]. The comparative results are shown in Table III. It is observed that the proposed scheme outperformed both machine learning techniques and SWNet.

The only relaxation what we have observed in this approach, there may be sentences with multiple bigram or trigram phrases. In that case, we calculate sentiment score of the testing sentence based on the negative and positive phrase. If the sentiment score is zero, then the proposed system is not able to classify the sentiment of that sentence. It simply classify as ambiguous.

TABLE III: Accuracy Comparison of Proposed Approach with Existing ML Techniques

Classification Method	**Binary**	**Ternary**
Using Classifiers	**Accuracy (%)**	**Accuracy (%)**
SVM	52	41.5
LR	88	77
NB	64	54
RF	86.5	74.6
MLP-NN	52	65
DT	84	76.1
Using SWNet		78.3
Using Proposed SPNet		**85.6**

V. Conclusion & Future Work

In the absence of sufficient annotated data-sets in the Telugu language, this research has focused to build a Telugu SPNet to identify the sentiment in Telugu data. Further, this work also covers for sentiment analysis in other Indian languages. The performance of SPNet is compared with existing machine learning approaches for sentiment analysis and existing SWNet approach for sentiment analysis in Telugu data. It is also shows that SPNet is performing better than existing Telugu SWNet and machine learning techniques. In future, we wish to apply dynamism to include more Phrases in SentiPhraseNet and also extends for other Indian languages.

Acknowledgment

The authors would like to thank Rongala Srikanth, Pelluru Pavan and Amarlapudi Mahesh Babu for the contribution while collecting dataset to build this SentiPhraseNet and also would like to thank Killi Bala Prakash, Bojanki Manikanta, Chintala Vijay and Vaddi Madhusudan for providing annotation of our collected data set. All the annotators belong to the state of Andhra Pradesh and their mother tongue is Telugu.

References

[1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[2] J. M. Wiebe, "Tracking point of view in narrative," *Computational Linguistics*, vol. 20, no. 2, pp. 233–287, 1994.

[3] TeluguRank. Ethnologue list. [Online]. Available: https://www.ethnologue.com/statistics/size

[4] Census. (2001) States of india by telugu speakers. [Online]. Available: https://en.wikipedia.org/wiki/States of India by Telugu speakers

[5] A. Das and S. Bandyopadhyay, "Sentiwordnet for indian languages," *Asian Federation for Natural Language Processing, China*, pp. 56–63, 2010.

[6] X. Wan, "Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 553–561.

[7] P. Aliandu, "Sentiment analysis on indonesian tweet," *The Proceedings of The 7th ICTS*, 2014.

[8] A. Das and S. Bandyopadhay, "Dr sentiment creates sentiwordnet (s) for indian languages involving internet population," in *Proceedings of Indo-wordnet workshop*, 2010.
[9] Y. Chen and S. Skiena, "Building sentiment lexicons for all major languages." in *ACL (2)*, 2014, pp. 383–389.
[10] 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, 2015, pp. 656–663.
[11] S. S. Kumar, B. Premjith, M. A. Kumar, and K. Soman, "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, 2015, pp. 671–683.
[12] K. Sarkar and S. Chakraborty, "A sentiment analysis system for indian language tweets," in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2015, pp. 694–702.
[13] B. G. 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, 2015, pp. 650–655.
[14] M. Venugopalan and D. 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, 2015, pp. 664–670.
[15] K. Ravi and V. Ravi, "Sentiment classification of hinglish text," in *Recent Advances in Information Technology (RAIT), 2016 3rd International Conference on*. IEEE, 2016, pp. 641–645.
[16] D. S. Nair, J. P. Jayan, E. Sherly *et al.*, "Sentima-sentiment extraction for malayalam," in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. IEEE, 2014, pp. 1719–1723.
[17] S. K. Sahu, P. Behera, D. Mohapatra, and R. C. Balabantaray, "Information retrieval in web for an indian language: An odia language sentimental analysis context," pp. 249–256, 2016.
[18] S. S. Mukku, I. LTRC, N. Choudhary, and R. Mamidi, "Enhanced sentiment classification of telugu text using ml techniques," in *25th International Joint Conference on Artificial Intelligence*, 2016, p. 29.
[19] R. Naidu, S. K. Bharti, K. S. Babu, and R. K. Mohapatra, "Sentiment analysis using telugu sentiwordnet," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017, pp. 666–670.
[20] S. S. Mukku, I. LTRC, and R. Mamidi, "Actsa: Annotated corpus for telugu sentiment analysis," in *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, 2017, pp. 54–58.
[21] S. Reddy and S. Sharoff, "Cross language pos taggers (and other tools) for indian languages: An experiment with kannada using telugu resources," in *Proceedings of IJCNLP workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies.*, Chiang Mai, Thailand, 2011.
[22] A. Bharati and P. R. Mannem, "Introduction to shallow parsing contest on south asian languages," in *Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL)*. Citeseer, 2007, pp. 1–8.
[23] J. L. Fleiss, J. Cohen, and B. Everitt, "Large sample standard errors of kappa and weighted kappa." *Psychological Bulletin*, vol. 72, no. 5, p. 323, 1969.