BCSAT : A Benchmark Corpus for Sentiment Analysis in Telugu Using Word-level Annotations

Sreekavitha Parupalli, Vijjini Anvesh Rao and Radhika Mamidi
Language Technologies Research Center (LTRC)
International Institute of Information Technology, Hyderabad
{sreekavitha.parupalli, vijjinianvesh.rao}@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

The presented work aims at generating a systematically annotated corpus that can support the enhancement of sentiment analysis tasks in Telugu using word-level sentiment annotations. From OntoSenseNet, we extracted 11,000 adjectives, 253 adverbs, 8483 verbs and sentiment annotation is being done by language experts. We discuss the methodology followed for the polarity annotations and validate the developed resource. This work aims at developing a benchmark corpus, as an extension to SentiWordNet, and baseline accuracy for a model where lexeme annotations are applied for sentiment predictions. The fundamental aim of this paper is to validate and study the possibility of utilizing machine learning algorithms, word-level sentiment annotations in the task of automated sentiment identification. Furthermore, accuracy is improved by annotating the bi-grams extracted from the target corpus.

1 Introduction

Sentiment analysis deals with the task of determining the polarity of text. To distinguish positive and negative opinions in simple texts such as reviews, blogs, and news articles, sentiment analysis (or opinion mining) is used. Over time, it evolved from focusing on explicit opinion expressions to addressing a type of opinion inference which is a result of opinions expressed towards events having positive or negative effects on entities.

There are three ways in which one can perform sentiment analysis : document-level, sentence-level, entity or word-level. These determine the polarity value considering the whole document, sentence-wise polarity, word-wise in some given text respectively (Naidu et al., 2017). Despite extensive research, the existing solutions and systems have a lot of scope for improvement, to meet the standards of the end users. The main problem arises while cataloging the possibly infinite set of conceptual rules that operate behind the analyzing the hidden polarity of the text (Das and Bandyopadhyay, 2011). In this paper, we perform a word-level sentiment annotation to validate the usage of such techniques for improving sentiment analysis task. Furthermore, we use word embeddings of the word-level sentiment annotated lexicon to predict the sentiment label of a document. We experiment with various machine learning algorithms to analyze the affect of word-level sentiment annotations on (document-level) sentiment analysis.

The paper is organized as follows. In section 2 we discuss the previous works in the field of sentiment analysis, existing resources for Telugu and specific advances that are made in Telugu. Section 3 describes our corpus and annotation scheme. 4 section describes several experiments that are carried out and the accuracies obtained. We also explain the results in detail in 4.3. Section 5 showcases our conclusions and section 6 shows the scope for future work.

2 Related Work

- `Sentiment Analysis:` Several approaches have been proposed to capture the sentiment in the text where each approach addresses the issue at different levels of granularity. Some researchers have proposed methods for document-level sentiment classification (Pang et al., 2002; Turney and Littman, 2003). At the top level of granularity, it is

Proceedings of ACL 2018, Student Research Workshop, pages 99–104
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

often impossible to infer the sentiment expressed about any particular entity, because a document may convey different opinions for different entities. Hence, when we consider the tasks of opinion mining where the sole aim is to capture the sentiment polarities about entities, such as products in product reviews, it has been shown that sentence-level and phrase-level analysis lead to a performance gain (Wilson et al., 2005; Choi and Wiebe, 2014). In the context of Indian languages, (Das et al., 2012) proposes an alternate way to build the resources for multilingual affect analysis where translations into Telugu are done using WordNet.

- `SentiWordNet :` (Das and Bandyopadhyay, 2010) proposes multiple computational techniques like, WordNet based, dictionary based, corpus based and generative approaches to generate Telugu SentiWordNet. (Das and Bandyopadhyay, 2011) proposes a tool Dr Sentiment where it automatically creates the PsychoSentiWordNet which is an extension of SentiWordNet that presently holds human psychological knowledge on a few aspects along with sentiment knowledge.

- `Advances in Telugu:` (Naidu et al., 2017) utilizes Telugu SentiWordNet on the news corpus to perform the task of Sentiment Analysis. (Mukku and Mamidi, 2017) developed a polarity annotated corpus where positive, negative, neutral polarities are assigned to 5410 sentences in the corpus collected from several sources. They developed a gold standard annotated corpus of Telugu sentences aimed at improving sentiment analysis in Telugu. To minimize the dependence of machine learning(ML) approaches for sentiment analysis on abundance of corpus, this paper proposes a novel method to learn representations of resource-poor languages by training them jointly with resource-rich languages using a siamese network (Choudhary et al., 2018a). A novel approach to classify sentences into their corresponding sentiment using contrastive learning is proposed by (Choudhary et al., 2018b) which utilizes the shared parameters of siamese networks.

  (Gangula and Mamidi, 2018) and (Mukku and Mamidi, 2017) are the only reported works for Telugu sentiment analysis using sentence-level annotations who developed annotated corpora. Ours is the first of it's kind NLP research which uses sentiment annotation of bi-grams for sentiment analysis (opinion mining).

3 Building the Benchmark Corpus

Lexicons play an important role in sentiment analysis. Having annotated lexicon is key to carry out sentiment analysis efficiently. The primary task in sentiment analysis is to identify the polarity of text in any given document. The polarity may be either positive, negative or neutral (Naidu et al., 2017). Sentiment is a property of human intelligence and is not entirely based on the features of a language. Thus, peoples involvement is required to capture the sentiment (Das and Bandyopadhyay, 2011). Having said this, we establish that annotated lexicons are of immense importance in any language for sentiment analysis (a.k.a opinion mining).

For our experiments, we utilize the reviews dataset from Sentiraama [1] corpus. It contains 668 reviews in total for 267 movies, 201 products and 200 books. Product reviews has 101 positive and 100 negative entries; movie reviews has 136 positive and 132 negative reviews; book reviews data has 100 positive and 100 negative entries. Since the obtained corpus is only annotated with document-level sentiment labels, we perform the word-level sentiment annotation manually.

3.1 Annotation Procedure

In this paper, sentiment polarities are classified into 4 labels : positive, negative, neutral and ambiguous. Positive and negative labels are given in case of positive and negative sentiments in the word respectively. Ambiguous label is given to words which acquire sentiment based on the words it is used along with or it's position in a sentence. Neutral label is given when the word has no sentiment in it. However, neutral and ambiguous sentiment labels are of no significant use for the task of sentiment analysis. Henceforth, those labels are ignored in our experiments.

Sentiment annotations are performed on two different kinds of data. Table 1 showcases the distribution of sentiment labels at the word-level.

[1] `https://ltrc.iiit.ac.in/showfile.php?filename=downloads/sentiraama/`

- `Unigrams`: We obtain 7,663 words from Telugu SentiWordNet [2] resource to calculate the base-line accuracy of any word-level sentiment annotated model. These words are already annotated for sentiment/polarity. However, it doesn't provide extensive coverage of Telugu. Later on, we discover a newly developed large resource of Telugu words by (Parupalli and Singh, 2018), OntoSenseNet, which has a collection of 21,000 words (adjectives+verbs+adverbs). We perform the task of word-level sentiment annotation on the words obtained from this resource and we refer to these annotated words as unigrams throughout this paper. Language experts who performed the annotations are given some guidelines to follow. Experts are implored to look at the word, it's gloss and then decide which one of the four sentiment labels is more apt for a given word. Aforementioned word-level sentiment annotation is an attempt to improve the coverage of SentiWordNet.

- `Bigrams`: Furthermore, sentiment cannot always be captured in a single word.This paper aims to check if bigram annotation is a suitable approach for improving the efficiency of sentiment analysis. To validate the hypothesis, we extract bigrams, which occurred at least more than once, only from the target corpus - Sentiraama dataset developed by (Gangula and Mamidi, 2018). For example, consider the bigram ('DhokA', 'ledu'). The words individually mean 'hurdle (DhokA)', 'no (ledu)'. Thus, in word-level annotation task they would be given a negative label. However, the bigram means there is 'nothing that can stop' which invokes a positive sentiment. Such occurrences are quite common in the text, especially reviews, which lead us to believe that bigram polarity has potential to enhance sentiment analysis, opinion mining. The usage of this developed resource in experiments performed is explained in section 4.

3.2 Validation

Annotations are done by 2 native speakers of Telugu. If the annotators aren't able to decide which label to assign, they are advised to tag it as uncertain. In case of a disagreement, the label given by the annotator with more experience is given priority. Validation of the developed resource is done using Cohen's Kappa (Cohen, 1968). By considering the uncertain cases as borderline cases (where at least one annotator tagged the word as uncertain), Kappa value is seen as 0.91. This shows almost perfect agreement and this proves the consistency in annotation task. This is especially high because when both the annotators are uncertain, we did a re-iteration to finalize the tag. Such reiterative task is done for about 2,400 words during the development of our resource.

4 Experiments and Results

In this section we will analyze and observe how word-level polarity affects overall sentiment of the text through majority polling approach and machine learning based classification approaches.

4.1 Majority Polling Approach

A simple intuitive approach to identify the sentiment label of the text is to calculate the sum of positive(+1) and negative(-1) polarity values in it. If the sum is positive, it shows that number of positive words have outnumbered the number of negative words thus resulting in a positive sentiment on the whole. Otherwise, the polarity of the text is negative. Cases where the sum equals to 0 are ignored. Following are the word-level polarities we consider for positive and negative labels:

- `Unigram`: We use the annotated unigram data that is discussed in 3. For each review, we consider the unigram labels to carry the majority polling approach.

- `Bigram`: The extracted bigrams are annotated for positive and negative polarity. Initially, we divide our data into training and testing sets in 7:3 ratio. We only consider the annotated bigrams from the training corpus to predict the sentiment polarity of reviews in the test data.

- `Unigram+Bigram`: In this trial, we combine the unigram and bigram data to perform majority polling. We consider the whole unigram data whereas bigrams extracted from the training set are only considered for predictions.

[2] `http://amitavadas.com/sentiwordnet.php`

Resource	Positive	Negative	Neutral	Ambiguous	Total
SentiWordNet [3]	2135	4076	359	1093	7663
Dictionary (Parupalli and Singh, 2018)	3080	4232	3391	10199	20896
Bigrams	1978	1762	8990	1996	14826

Table 1: Distribution of Sentiment Labels in Several Resources

Furthermore, as Telugu is agglutinative in nature (Pingali and Varma, 2006), we experiment with the above mentioned approaches after performing morphological segmentation provided by Indic NLP library [4]. Morphological segmentation is performed on the original reviews data and n-grams (positive and negative labels) to see if we could get more accurate sentiment prediction of the reviews due to increment in the coverage.

4.2 Machine Learning Based Classification Approach

In this section, we perform document-level sentiment analysis task with word embedding models, specifically Word2Vec. We utilize a Word2Vec model that is trained on corpus consisting of scrapped data from Telugu websites, with 270 million non-unique tokens on the whole. Furthermore, to obtain vectors for each review, we take word vector of every word in the review and calculate their average to get a single document vector.

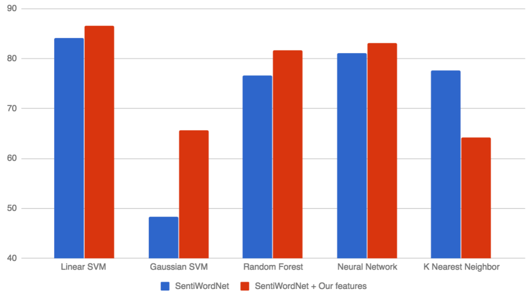

Figure 1: Comparative analysis of percentage accuracies produced by various classifiers

Though traditional vector-based word representations help us accomplish various natural language processing tasks, they often lack information related to sentiment analysis. Thus, we aim to enrich the Word2Vec vectors obtained from corpus by incorporating word-level polarity features. We do this by adding the features we propose in 4.1 to the original averaged Word2Vec vector, which is expected to increase the accuracy of polarity prediction. The additional features we added are: *positive unigrams* (number of positive polarity unigrams in the review), *negative unigrams* (number of negative polarity unigrams in the review), *positive bigrams* (number of positive polarity bigrams in the review), *negative bigrams* (number of negative polarity bigrams in the review). We partition these document vectors into training and testing sets to develop various classifier models. In this paper, we have implemented 5 classifiers, namely, Linear SVM, Gaussian SVM, Random Forest, Neural Network, K Nearest Neighbor (KNN). Percentage accuracies are illustrated along with the improvement in accuracies after addition of our proposed features in Figure 1 and results are discussed in 4.3.

4.3 Results

In this section, we showcase and analyze the results of the two experiments we have done in section 4.

4.3.1 Majority Polling Approach :

Results illustrated in Table 2 show that certain word-level features do capture information relevant to document-level sentiment analysis. Our hypothesis in Section 3.1 shows that bigram polarity annotations have potential to enhance sentiment analysis. High accuracy obtained by using only bigrams for majority polling proves our hypothesis. However, there is a trade-off between coverage and accuracy. This can be depicted from the huge increase in the count of unclassified reviews in case of bigram majority poling. We also observe that effect of morphological segmentation on accuracy is hardly positive. This indicates that in case of Telugu, morphological data has relevance to sentiment expressed and morphological segmentation would result in loss of such valuable information for sentiment analysis tasks.

[4] http://anoopkunchukuttan.github.io/indic_nlp_library/

	SentiWordNet	Our resource	Bigram	Uni+Bigrams
Before Segmentation	61.86	62.84	78.97	55.44
Unclassified reviews	23/201	14/201	108/201	10/201
After Segmentation	60.23	58.29	49.46	57.89
Unclassified reviews	20/201	18/201	36/201	8/201

Table 2: Comparison of accuracies obtained through majority polling on different resources.

4.3.2 Machine Learning Based Classification Approach:

This approach shows that across all the classifiers, addition of word-level polarity features improves the process of classification. Therefore, classifiers can predict document-level sentiment polarity with better accuracies. Hence, our hypothesis is validated once again. Accuracies doesn't improve significantly over the baseline value but show a small increment always. KNN classifier shows a huge drop in accuracy after inclusion of the new features proposed. This is observed because KNN assumes all features to hold equal importance for classification. Hence, KNN fails to ignore the noisy features which explains the drop. Random forest and neural network classifiers don't show significant learning from the proposed features. Finally, we observe that linear SVM classifier works best to identify the polarity of a text for our features indicating linear separability of the data. This also explains the bad performance of Gaussian SVM. Linear SVM produces an accuracy of 84.08% when SentiWordNet words alone are used as a feature, which can be considered as the baseline accuracy. It gives an accuracy of 83.44% ,84.34% and 86.57% for unigrams, bigrams and unigrams+bigrams respectively as features of Linear SVM classifier.

5 Conclusions

In this paper, efforts are made to develop an annotated corpus of 21,000 words to enrich Telugu SentiWordNet. This is a work in progress. We perform annotations of 14,000 bigrams that are extracted from target corpus to validate their importance. This is a first-of-it's-kind approach in Telugu to enhance sentiment analysis. Manual annotations done show perfect agreement which validates the developed resource. Furthermore, we provide a justification to why word-level sentiment annotation of bigrams enhances sentiment analysis though an intuitive majority polling approach, by using several ML classifiers. The results are analyzed for further insights.

6 Future Work

We extract bigrams only from the target corpus because we wanted to mainly validate the importance of bigrams in sentiment analysis. However, attempts should be made to enhance the SentiWordNet with, at least, some most occurring bigrams in Telugu. We hope this corpus can serve as a basis for more work to be done in the area of sentiment analysis for Telugu. A continuation to this paper could be handling the enrichment of adjectives and adverbs available in OntoSenseNet for Telugu.

6.1 Crowd sourcing

We can develop a crowd sourcing platform where the annotations can be done by several language experts instead of a few. This helps in the annotation of large corpora. We aim to develop a crowd sourcing model for the same in near future. This would be of immense help in annotation of 21,000 unigrams extracted from the dictionary developed by (Parupalli and Singh, 2018).

7 Acknowledgements

This work is part of the ongoing MS thesis in Exact Humanities under the guidance of Prof. Radhika Mamidi. I am immensely grateful to Vijaya Lakshmi for helping me with data collection. I would like to thank Nurendra Choudary for reviewing the paper and for his part in the ideation. I would like to extend my gratitude to Abhilash Reddy for annotating the dataset, reviewing the work carefully and constantly pushing us to do better. I want to thank Rama Rohit Reddy for his support and for validating the novelty of this research at several points. I acknowledge the support of Google in the form of an International Travel Grant, which enabled me to attend this conference.

References

Yoonjung Choi and Janyce Wiebe. 2014. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191.

Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018a. Emotions are universal: Learning sentiment based representations of resource-poor languages using siamese networks. *arXiv preprint arXiv:1804.00805*.

Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018b. Sentiment analysis of code-mixed languages leveraging resource rich languages. *arXiv preprint arXiv:1804.00806*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Amitava Das and Sivaji Bandyopadhyay. 2010. Sentiwordnet for indian languages. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 56–63.

Amitava Das and Sivaji Bandyopadhyay. 2011. Dr sentiment knows everything! In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: systems demonstrations*, pages 50–55. Association for Computational Linguistics.

Dipankar Das, Soujanya Poria, Chandra Mohan Dasari, and Sivaji Bandyopadhyay. 2012. Building resources for multilingual affect analysis–a case study on hindi, bengali and telugu. In *Workshop Programme*, page 54.

Rama Rohit Reddy Gangula and Radhika Mamidi. 2018. Resource creation towards automated sentiment analysis in telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Sandeep Sricharan Mukku and Radhika Mamidi. 2017. Actsa: Annotated corpus for telugu sentiment analysis. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 54–58.

Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu, and Ramesh Kumar Mohapatra. 2017. Sentiment analysis using telugu sentiwordnet.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Sreekavitha Parupalli and Navjyoti Singh. 2018. Enrichment of ontosensenet: Adding a sense-annotated telugu lexicon. *arXiv preprint arXiv:1804.02186*.

Prasad Pingali and Vasudeva Varma. 2006. Hindi and telugu to english cross language information retrieval at clef 2006. In *CLEF (Working Notes)*.

Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.