ACTSA: Annotated Corpus for Telugu Sentiment Analysis

Sandeep Sricharan Mukku and **Radhika Mamidi**
Language Technologies Research Center
KCIS, IIIT Hyderabad
sandeep.mukku@research.iiit.ac.in, radhika.mamidi@iiit.ac.in

Abstract

Sentiment analysis deals with the task of determining the polarity of a document or sentence and has received a lot of attention in recent years for the English language. With the rapid growth of social media these days, a lot of data is available in regional languages besides English. Telugu is one such regional language with abundant data available in social media, but it's hard to find a labelled data of sentences for Telugu Sentiment Analysis. In this paper, we describe an effort to build a gold-standard annotated corpus of Telugu sentences to support Telugu Sentiment Analysis. The corpus, named ACTSA (Annotated Corpus for Telugu Sentiment Analysis) has a collection of Telugu sentences taken from different sources which were then preprocessed and manually annotated by native Telugu speakers using our annotation guidelines. In total, we have annotated 5410 sentences, which makes our corpus the largest resource currently available. The corpus and annotation guidelines are made publicly available.

1 Introduction

Now-a-days, people are commonly found writing comments, reviews, blog posts in social media about trending activities in their regional languages. Unlike English, many regional languages lack NLP tools and resources to analyze these activities. Moreover, English has many datasets available, however, it is not the same with Telugu.

The annotation of Telugu data has not received a lot of attention in sentiment analysis community. While there is a wealth of raw corpora with opinionated information, no corpora with annotated sentences in Telugu are publicly available as far as we know.

Telugu has a special status as an official standard language in the twin states of Andhra Pradesh and Telangana of India. There are a large variety of dialects that constitute the mother tongues of Telugu speakers. Major Telugu print media, journalism, and electronic media follow the dialects of Krishna and Godavari since it has been conceived as arguably standard and easy to reach the rest of the Telugu speakers (Krishnamurthi, 1961). We built our corpus over this dialect as this dialect is most prominent and has a strong online presence today on news websites, blogs, forums, and user/reader commentaries.

In this work, we present a dedicated gold standard corpus of polarity annotated Telugu sentences. To our knowledge, our corpus is the largest source of polarity annotated Telugu sentences to date. This data also motivates the development of new techniques for Telugu sentiment analysis. The corpus and annotation guidelines are publicly available here[1].

2 Related Work

There is a growing interest within the Natural Language Processing community to build corpora for Indian languages from the data available on the web. (Kaur and Gupta, 2013) surveyed sentiment analysis for different Indian languages including Telugu, but never mentioned about the corpus used. (Mukku et al., 2016) did sentiment classification for Telugu

[1] https://goo.gl/M9rkUX

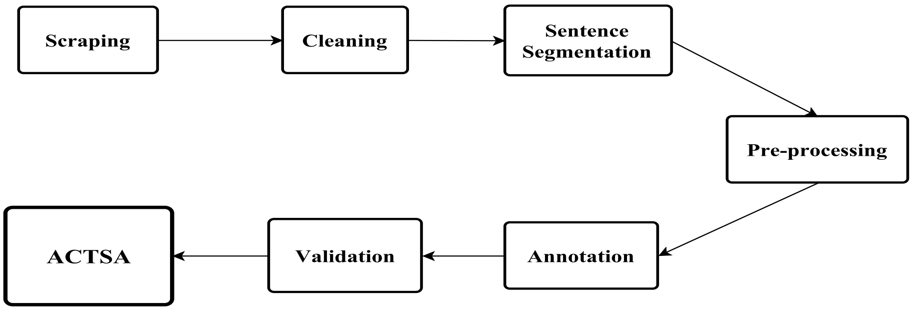

Figure 1: Process of building the resource

text using various ML techniques, but no data was publicly made available.

(Wiebe et al., 2005) describes a corpus annotation project to study issues in the manual annotation of opinions, emotions, sentiments, speculations, evaluations and other private states in language. This was the first attempt to manually annotate the 10,000 sentence corpus of articles from the news. (Alm et al., 2005) have manually annotated 1580 sentences extracted from 22 Grimms' tales for the task of emotion annotation at the sentence level.

(Arora, 2013) performed sentiment analysis task for the Hindi Language with limited corpus made manually annotated by the native Hindi speakers. (Das and Bandyopadhyay, 2010b) aims to manually annotate the sentences in a web-based Bengali blog corpus with the emotional components such as emotional expression (word/phrase), intensity, associated holder and topic(s).

(Das and Bandyopadhyay, 2010a) built a lexicon of words to support the task of Telugu sentiment analysis and is made available to the public. (Das and Bandyopadhay, 2010) created an interactive gaming to technology (Dr. Sentiment) to create and validate SentiWordNet for Telugu.

3 Data Collection

In this section, we will explore the different resources where raw data was obtained from and how processing of that data was done, as shown in Figure 1.

Currently, most of the corpora available for Sentiment Analysis are harvested from sources like review data from e-commerce websites where customers express their opinion on products freely, posts from social networking sites like Twitter and Facebook. Although the news genre has received much less attention within the Sentiment Analysis community, news plays an important role in exhibiting the reality and has a strong influence on social practices. Also, a lot of Telugu data is available mostly on news websites. These reasons motivated us to select news genre for building our corpus. We scraped and harvested our raw data from five different Telugu news websites viz., Andhrabhoomi[2], Andhrajyothi[3], Eenadu[4], Kridajyothi[5] and Sakshi[6]. In total we have collected over 453 news articles and filtered down to 321 which were relevant to our work.

The extracted data was cleaned in a preprocessing step, e.g. by removing headings and sub-headings, eliminating sentences with non-Telugu words and cleaning any extra dots, extra spaces, URLs, and other garbage values. Later *Sentence Segmentation* is done where this data was split into individual sentences.

The sentences thus obtained were now tested for objectivity manually. Objective sentences are sentences where no sentiment, opinion, etc. is expressed. They state a fact confidently and has an evidence to support it. For example, sentence (1) is an objective sentence as it is a verifiable fact with evidence.

అబ్దుల్ కలాం భారతదేశ అధ్యక్షుడిగా పనిచేశారు (1)

Transliteration: Abdul kalāṁ bhāratadēśa adhyakṣuḍigā panicēśāru
English: Abdul Kalam served as the president of India

[2] http://www.andhrabhoomi.net/
[3] http://www.andhrajyothy.com/
[4] http://www.eenadu.net/
[5] http://www.andhrajyothy.com/pages/sports
[6] http://www.sakshi.com/

Table 1: Example annotations

ID	Original Sentence	English Translation	A1	A2	V	F
1	అమెరికా అధ్యక్షుడు డోనాల్డ్ ట్రంప్ పారిస్ వాతావరణ ఒప్పందం నుంచి అమెరికాను వైదొలగించారు	US President Donald Trump withdrew the US from the Paris Climate Agreement	Neg	Obj	Obj	Obj
2	ఇందుకు ఎవరికీ అభ్యంతరం ఉండనవసరం లేదు	There is no need for any objection to anyone in this	Neu	Neu	NA	Neu
3	భారత ప్రధానమంత్రి నరేంద్రమోడీ కాశ్మీర్ అల్లర్లపై ఘాటుగా స్పందించారు	India's Prime Minister Narendra Modi has reacted severely to the Kashmir riots	Neg	Neg	NA	Neg
4	ఫలితాలపై మంత్రి సంతోషంగా ఉన్నారు	The minister is happy on the results	Pos	Pos	NA	Pos

Pos = Positive, Neg = Negative, Neu = Neutral, Obj = Objective, NA = Not Applicable, A1 = Annotator 1, A2 = Annotator 2, V = Validation, F = Final Result

These sentences do not contain any sentiment/polarity and are not useful for sentiment analysis. The objective sentences thus separated with objectivity test are removed from the data.

4 Annotation

In this section, we describe the process followed for annotating the sentences (refer Figure 1). First, we built a team of seven educated native Telugu speakers for the task of polarity tagging of the extracted Telugu sentences. Then, we developed an annotation schema for this task and the annotators were instructed to thoroughly understand the concepts mentioned in the schema for a precise/perfect annotation. Each sentence is annotated by two annotators.

The annotators were required to tag the sentences with three polarities: *positive, negative, neutral.* For example, sentence (2) should be tagged *positive* as it expresses positive sentiment by the use of కృతజ్ఞత (gratitude).

మంత్రి, ఆయనను ఎన్నుకున్నందుకు, (2)

ప్రజలకు కృతజ్ఞత వ్యక్తం చేశారు
Transliteration: Mantri, āyananu ennukunnanduku, prajalaku krtajñata vyaktaṁ cēśāru
English: The minister expressed gratitude to the people for electing him

On the other hand sentence (3) should be tagged *negative* because it expresses negative sentiment with ఆందోళన (concern).

నిరంతర విద్యుత్ కోతలపై ప్రజలు ఆందోళన వ్యక్తం చేశారు (3)

Transliteration: Nirantara vidyut kōtalapai prajalu āndōḷana vyaktaṁ cēśāru
English: People have expressed concern over continuous power cuts

However, sentence (4) is a *neutral* sentence as it is a speculation about the future. Even though it doesn't contain any sentiment, it is not an objective sentence because it is not a verifiable fact or not something which happened in the past. It is speculating something to happen in the future.

ప్రధాని వచ్చే నెలలో చైనాను సందర్శించనున్నారు (4)

Transliteration: Pradhāni vaccē nelalō cainānu sandarśiñcanunnāru
English: The prime minister is expected to visit China next month

If in any case annotators were unsure or felt ambiguous about the polarity of a sentence they can label it *uncertain.* If they feel the sentence is objective but was not removed in

Table 2: Agreement for Sentences in ACTSA

Annotator 2 Annotator 1	Positive	Negative	Neutral	Total
Positive	1463	31	103	1597
Negative	23	1421	116	1560
Neutral	112	127	2427	2666
Total	1598	1579	2646	5823

the pre-processing step, they can mark it *objective*.

Annotators labelled all the sentences, with each sentence annotated by exactly two annotators. We call it *annotation* step. The sentences marked *uncertain* by at least one annotator were discarded to avoid any ambiguous sentences in the corpus.

The sentences which had a clash between the two annotators' labels were sent for a *third independent annotation* which we call as a *validation* step. The most common label among the three annotators was considered as the final label for the sentence. If even after the third annotation the disagreement prevailed, such sentences were discarded as we considered them too ambiguous for getting three different labels by three different annotators. If there were any objective sentences after the *validation* step, they were discarded.

Table 1 shows some example annotations from the corpus.

5 Agreement Study

After annotation task, we measured how reliable our annotation scheme was. To measure the reliability of our polarity annotation scheme, we conducted an inter-annotator agreement study on the annotated sentences. Table 2 shows the agreement for the two annotators' judgments for each sentence. We used Cohens´ kappa, κ which is calculated using formula (5)

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

where p_o is the relative observed agreement and p_e is the agreement by chance. In general, κ values between 0.6 and 0.8 are considered a substantial agreement. To our surprise we got the κ value to be 0.87, which is in perfect agreement and is an indication of the reliability of the annotations.

Table 3: Statistics about the data

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

6 Corpus Statistics

In this section, we present the statistics about our data from raw data collection to final sentences. We scraped several websites for the data. We collected 453 news articles and filtered down to 321 which were relevant for our work. After pre-processing this raw data, we have 11952 sentences. We tested the sentences for subjectivity (as explained in section 3) and removed 4327 objective sentences after which we were left with 7812 sentences. These sentences were given to the annotators for the annotation as mentioned in section 4. 1802 sentences were removed where at least one annotator marked it *uncertain*. In the remaining 5823 sentences, 512 were with disagreement and were sent for third independent annotation. After the third annotation, 413 sentences were discarded if the disagreement prevailed or if they are objective. The final 5410 sentences forms the required annotated corpus, ACTSA. Statistics about our complete corpus can be found in Table 3.

7 Experiments and Evaluation

A strategy that can give very useful hints about the reliability of the annotated data is the comparison between the results of automated classification and human annotation.

(Mukku et al., 2016) described a method to perform automated classification of Telugu sentences into polarity tags: positive, negative and neutral. We followed this method to evaluate our data. We used 2000 sentences from our human automated corpus to train the model for automated classification.

To test the reliability of our annotated data, we compared the classification expressed by humans and that of the automated classifier trained above. The testing was done on the remaining 3410 sentences and the error rate was observed to be **12.3%** which hints the quality and reliability of the annotated corpus, ACTSA.

8 Conclusion

In this work, we presented a gold standard corpus of Telugu sentences taken from different resources, which were then cleaned and annotated by native Telugu speakers. For each sentence, we have a polarity label attached with it. We described our annotation process and gave an overview of our annotation schema. The results from our evaluation study show that our corpus has a reasonable inter-annotator agreement. The corpus and guidelines are publicly available. In future, we try to automate the task of annotation for new sentences with the help of ACTSA. We would also like to perform sentiment analysis task for Telugu, using this corpus.

Acknowledgements

We would like to thank all the members of annotation group (native Telugu speakers) for their tireless effort in this process. The work reported in this paper was supported by Kohli Center on Intelligent Systems.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 579–586.

Piyush Arora. 2013. Sentiment analysis for hindi language. *MS by Research in Computer Science, IIIT Hyderabad* .

Amitava Das and S Bandyopadhay. 2010. Dr sentiment creates sentiwordnet (s) for indian languages involving internet population. In *Proceedings of Indo-wordnet workshop*.

Amitava Das and Sivaji Bandyopadhyay. 2010a. Sentiwordnet for indian languages. *Asian Federation for Natural Language Processing, China* pages 56–63.

Dipankar Das and Sivaji Bandyopadhyay. 2010b. Labeling emotion in bengali blog corpus–a fine grained tagging at sentence level. In *Proceedings of the 8th Workshop on Asian Language Resources*. page 47.

Amandeep Kaur and Vishal Gupta. 2013. A survey on sentiment analysis and opinion mining techniques. *Journal of Emerging Technologies in Web Intelligence* 5(4):367–371.

Bh Krishnamurthi. 1961. Telugu verbal bases: A comparative and descriptive study.

Sandeep Sricharan Mukku, Nurendra Choudhary, and Radhika Mamidi. 2016. Enhanced sentiment classification of telugu text using ml techniques. In *4th Workshop on Sentiment Analysis where AI meets Psychology, 25th International Joint Conference on Artificial Intelligence*. page 29.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39(2):165–210.