Polarity Classification for Telugu Tweets

K. DAVID RAJU#1, DR. BIPIN BIHARI JAYASINGH#2, DR. P. VIJAYAPAL REDDY#3

#1Research Scholor, Rayalaseema University, KURNOOL,AP,

#2Professor, Dept of IT, CVR College of Engineering, Telangana

#3 Professor, Dept of CSE, Matrusri Engineering College, Telangana,

***Abstract*—**Social networking services such as Facebook and Twitter and social media hosting websites such as Flickr and YouTube have become increasingly popular in recent years. One key factor to their attractiveness worldwide is that these sites and services allow people to express and share their opinions, likes, and dislikes, freely and openly. The opinions posted range from criticizing politicians to discussing cricket matches, citing top news,criticizing political parties, appraising movies, and recommending new products and services such as mobiles, restaurants, and software. This development has fueled a new field known as sentiment analysis and opinion mining with the goal of extracting people's sentiment from text to assist customers in their purchase decisions and vendors in enhancing their reputation. This emerging field has attracted a large research interest, but most of the existing work focuses on English text. Hence, in this thesis, we studied sentiment analysis of telugu text retrieved from a well-known social media site, namely Twitter. Specifically, we studied the topic of target-dependent sentiment analysis of Telugu Twitter text, which has not been addressed in Telugu language before. We developed a system that will acquire Telugu text from Twitter and extract users opinions towards different topics and products.

Keywords*—Morphologically Rich Language, Subjectivity and Sentiment Analysis, Natural Language Processing, Telugu*

I. INTRODUCTION

Telugu is a Dravidian language, native to India. It ranks third by the number of native speakers in India and fifteenth in the Ethnologue list1 of the most spoken languages world-wide[2]. Over the last decade, there has been an increment in movie review sites, newspaper websites, tweets, comments and other blogposts etc., written in Telugu. Labeling these reviews with their sentiments would provide a brief summary to the readers. Unlike English, many regional languages lack resources to analyze these activities. Moreover,English has many datasets available, however, it is not the same with Telugu. Telugu has neither a large annotated dataset and tools nor any pre-trained models. Telugu data requires indispensable preprocessing for information extraction and sentiment extraction. Sentiment analysis hasattracted much research interest in the last decade, mostly in the English language. Researchers have analyzed sentiments in a variety of domains: movie reviews, news articles, blogs, forums, product reviews, and more recently social media data [1]. In the last few years, Telugu language sentiment analysis has started to attract some research interest as well.

II. SYSTEM DESCRIPTION

A. Model Generation

Generating a model for sentiment analysis is usually a lengthy pipelined process[10]. It consists of the six phases[3] presented in Figure 5.The first phase is typically Data Acquisition, followed by Tweet-Filtering phase, then Data Annotation phase, Data-Preprocessing phase, Feature Identification phase, and finally Classification Phase.

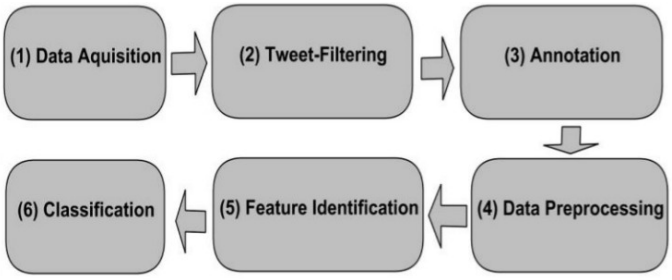


Figure 1: Model generation process

B. System Deployment

System deployment is also a sequential process that consists of four phases. The first phase is data acquisition, followed by data-preprocessing phase, and then feature identification phase, and finally classification phase. The process is presented in Figure 2.

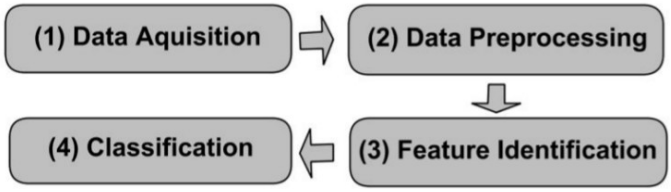


Figure 2: system deployment

III. DATA ANALYSIS

A. Annotation

In our system, we collected the data from Twitter.com. We have built a module to collect data from Twitter and we called it Twitter Fetcher. Twitter data and text has many peculiarities and language conventions such as the use of links, hash tag, emoticons, limited text length, etc. These peculiarities and conventions make it the nosiest social media site.

We collected data using the Twitter filter stream API using Twitter4j, an unofficial Java library for the Twitter API. The retweeted-field specifies whether this tweet has been retweeted or not.

We have chosen five keywords to tracks, pertaining to five different domains: technology, politics, news, sports and media respectively. Those keywords are: “Jio phone”(mobile device), “Modi” (India Prime Minister), “Special Status” (Protests for Special Status for Andhra Pradesh state), “Sunrisers” (IPL franchise cricket team) and “Jabardasth” (A popular Telugu TV show).To track any word we had to include several forms of the word.

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.

Transliteration:Abdul kalām bhāratadēśa adhyakṣuḍigā panicēśāru

English: Abdul Kalam served as the president of India

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.

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 three annotators and majority class is taken as true label. 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).

Transliteration: Mantri, āyananu ennukunnanduku, prajalaku krfftajñata vyaktam cēśāru

English: The minister expressed gratitude to the people for electing him

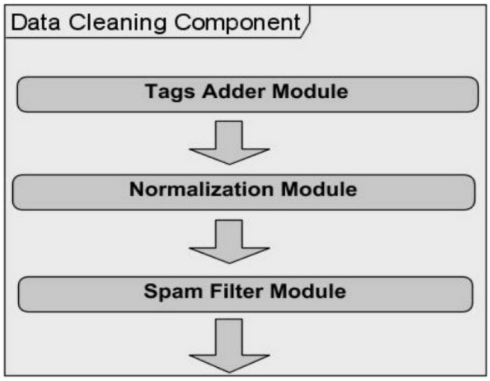
B. Data Preprocessing Phase

This phase comprises two steps as presented in Figure 11: In the first step, we performed data cleaning and in the second step, we performed various Natural Language Processing (NLP) tasks[2].This phase is crucial since it deals with the noisy nature of Twitter data.

The data-cleaning component consists of many modules: tags adder, normalization module, emoticons, laugh recognizer and spam filter (this phase is represented in Figure 12)

Table 1. Semantic attributes

Attribute Name	Definition
noPosWords	Number of positive words in the tweet. For example, “I love Arab Idol and adore the jury member” in this tweet there are 2 positive words namely “love”, and “adore”.
noNegWords	Number of negative words in the tweet. For example, “This restaurant food disgusts me I hate this restaurant” in this tweet we have two negative words namely “disgust”, and “hate”.
noPosEmo	Number of positive Emoticons. For example, “I’m soo happy :) ;)” has two positive emoticons.
noNegEmo	Number of Negative Emoticons. For example, “I’m soo sad : (;(” has two positive emoticons.
nolaugh	Number of laughters. For example, “hahaha did you see yesterday show @user it was funny loool” in this tweet we have 2 laughters namely “hahaha” and “loool”.
shortPos	The shortest distance between a positive word and the target of the sentiment in characters (incl. spaces).
shortNeg	The shortest distance between a negative word and the target of the sentiment in characters (incl. spaces).
nearPos	1 if the distance between the nearest positive word and the target word is shorter than the distance between the nearest negative word and the target word, 0 otherwise.
nearNeg	1 if the distance between the nearest negative word and the target word is shorter than the distance between the nearest positive word and the target word, 0 otherwise.



With various NLP Tasks[4], we tested the effectiveness of two Natural language processing tasks, stemming and Part-of-Speech tagging, on Telugu text sentiments analysis.

D. Feature Identification

This section describes the three types of features we extracted: syntactic features and semantic features[7]

- Syntactic features

  We used NB classifier with a presence vector where the features are the words N-grams and evaluated its performance using 10-fold cross- validation. The results showed that using a unigram only yields the best performance and therefore we carried out all the remaining experiments with unigrams. This result was expected because higher order N-grams leads to a very sparse feature space, which will not help the machine-learning algorithm in detecting pattern.

- Semantic features

  We have extracted nine semantic features, which are presented in Table 1 with their definitions. To extract these attributes, we had first to build a lexicon manually. Then, we extracted unigrams from those tweets and asked a human annotator to label those unigrams as positive or negative. We specifically excluded from the annotation process words that can be positive or negative depending on the context they are used in. For example, the word “fast” can be positive in this tweet, “Jio phone updates install very fast”, and it can be negative in this tweet “Jio phone battery runs out very fast”.

IV. NON-LEXICAL FEATURES EXTRACTION EXPERIMENTS

a) We used 17 non-lexical attributes[8], which include three syntactic features, namely: Number of punctuation marks, number of question marks, and number of exclamation marks . Nine semantic attributes (presented in Table 1) and five stylistic attributes which are: Number of usernames in tweet, Number of times the tweet was re-tweeted and Number of hash-tags in the tweet, Number of URL in the tweet, and Number of digits. Several experiments were conducted. In all the experiments, cleaned tweets were used (Data cleaning process is presented in section 3A, 3B) and the 17 non-lexical attributes were extracted from them. Polarity classification was performed to test the effectiveness of the 17 non-lexical attributes (classifying subjective tweets as positive or negative). The used datasets are: “Modi”, “Sunrisers”, “Jio phone” and “Special Status”. The combinations of these four datasets consist of 513 negative tweets and 612 positive tweets. For evaluation, 10-fold cross-validation was used.

The 17 attributes described are extracted and tested using four algorithms from Weka namely: SVM light (called LibSVM in Weka), SVM (called SMO in Weka), NB and J48 Decision Tree. SVM light is an implementation of Vapnik's Support Vector Machine [5] while SMO implements John Platt's [6] sequential minimal optimization algorithm for training a support vector classifier. The result is presented in Table 2. The table presents accuracy and Kappa statistic. Kappa statistic is a measure of reliability that indicates the proportion of agreement beyond that expected by chance. When there is no agreement other than that which would be expected by chance Kappa is zero. When there is a total agreement Kappa is one. In Weka, the Kappa statistic measures the agreement of prediction with the true class. We used our manually created polarity lexicon presented before to extract the semantic attributes. Initially, the lexicon had 890 positive words and 853 negative words. We have also created a Named Entity dictionary (seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, etc.) which included 59 words initially. The result shows that all classifiers perform similar to a random classifier.

TABLE 2.POLARITY CLASSIFICATION

The algorithm	Accuracy	Kappa Statistic
SVM (LibSVM)	60.5033 %	0.0745
SVM (SMO)	59.2889 %	0.1747
Naïve Bayes	55.4002 %	0.1035
J48 Tree	58.0217 %	0.1031

V. CONCLUSION

We developed a system named Telugu Subjectivity and Sentiment Analysis[9], which collects tweets from Twitter about specific topics. The topics in our experiments were "Modi", "Jio phone", "Sunrisers", "Special Status" and "Jabardasth". We called these topics the target of the sentiment and annotated the tweets towards the chosen target as "positive", "negative", or "neutral". We calledthis process target-dependent sentiment annotation. Our annotators were two native Telugu language speakers. After inner-annotator agreement, we were left with 4696 annotated tweets: 3287 tweets for training (70%) and 1409 for testing (30%). To develop our Telugu Subjectivity and Sentiment Analysis system, two stages were performed. In the first stage, we built a computational model that firstly identifies subjective text and then performs polarity classification. In the second stage, we employed the developed model to classify new tweet instances. The first stage consisted of six phases: Data Acquisition, followed by Tweet-Filtering phase, then Data Annotation phase, Data-Preprocessing phase, Feature Identification phase, and finally Classification Phase. The second stage consisted of four phases: Data Acquisition, Data-Preprocessing phase, Feature Identification phase, and finally Classification Phase.

We conducted several experiments to improve sentiment analysis accuracy. Initially, we tested the effect of data preprocessing on accuracy and we found significant improvement. Then, we examined the effect of several natural language processing tasks such as stemming and Part-of-Speech tagging but they did not prove useful. We extracted stylistic features, semantic features and syntactic features. Then, we tested the effect of the different features types on the sentiment accuracy. We can summarize our research contributions as follows:

1. We developed the first target dependent sentiment analysis system for Telugu language.

2. We experimented with different feature types, classification techniques, and classification algorithms to find which one suits best the given problem.

3. We identified several problems and issues regarding Telugu text (coming from Twitter) collection, annotation, and classification that have never been discussed in previous research.

4. We identified a set of rules that can be followed to facilitate sentiment classification.

5. We proved that manual inspection of Twitter text is necessary in order to identify issues and problems that would never have been identified otherwise.

Although our system was built specifically to analyze sentiment from tweet i.e., the text of messages from Twitter, it can be extended to handle other social media sites with few tweaks.For instance, Facebook doesn't present "mention" in the same way Twitter does. On Facebook, "mention" is just the user name while on Twitter the user name is preceded by the symbol "@". Thus, we must consider this difference in the Tag Adder module. In addition, Facebook does not support "hashtag" but this difference will not require any change in our system since the "HASHTAG" tag is added only when a hashtag is encountered. The same applies to the matter of "retweet". Moreover, Facebook status can be as long as 63206 characters long, which means we may have many sentences. Our system "as-is" can handle short single sentences, thus it doesn't need a co-reference resolution tool, but if we were to analyze a Facebook status then we would need such tool, as well as a sentence boundary detection tool. Furthermore, in the classification stage we would need to classify each sentence alone and then aggregate the results to find the overall sentiment towards a specific topic. In addition, a specific data fetcher must be developed for each social media site we wish to support. YouTube's comments and video descriptions can also be supported by our system. In fact, YouTube comments are very similar to Facebook status, as YouTube allows a maximum of 5000 characters for a playlist description.

In the future, if we obtain a fund, we are planning to run a study to test the effect of annotation process on sentiment analysis accuracy. The factors we are planning to study are: number of annotators, gender of annotators, language proficiency of annotators, age ofannotators, and degree of certainty of annotators regarding their annotation. In addition, we are planning to build large sentiment lexicon. To build this lexicon, we will collect large sample of tweets from different domains and about various topics. Then, we would use automatic and manual methods to annotate the lexicon words and phrases. Also, we are planning to annotate a large number of tweets after knowing the optimal annotation settings (number of annotators, age of annotator, etc.) and make this dataset public for research community.

REFERENCES

[1] B. Bosker, "Twitter Finally Shares Key Stats: 40 Percent Of Active Users Are Lurkers," 8 September 2011.

[2] M. Moussa, M. W. Fakhr and K. Darwish, "Statistical Denormalization for Telugu Text," in The Conference on Natural Language Processing ("Konferenz zur Verarbeitung Natürlicher Sprache", KONVENS, Vienna, 2012.

[3] G. Holmes, A. Donkin and I. H. Witten, "Weka: A Machine Learning Workbench," in Intelligent Information Systems, 1994.

[1] L. S. Larkey, L. Ballesteros and M. E. Connell, "Light Stemming for Telugu information Retrieval," Telugu Computational Morphology, pp. 221-243, 2007.

[2] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 2000.

[3] J. C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in Advances in Kernel Methods - Support Vector Learning, Washington,Microsoft Research, 1998.

[4] Y. Yang and J. O. Pedersen, "A Comparative Study On Features selection In Text Categorization," in International Conference on Machine Learning, Nashville, 1997.

[5] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data," in ACL Workshop on Language in Social Media (LSM), 2011.

[6] A. Joshi, A. Balamurali, P. Bhattacharyya and R. K. Mohanty, "C-Feel-It: A Sentiment Analyzer for Micro-Blogs," in Association for Computational Linguistics: Human Language Technologies, Portland, 2011.

[7] A. Go, R. Bhayani and L. Huang, "Twitter Sentiment Classification using Distant Supervision," Stanford Digital Library Technologies Project, 2009.

[8]