Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 1487-1493

http://www.ripublication.com

An Approach to Sentiment Analysis on Gujarati Tweets

Vrunda C. Joshi
PG Student, Computer Engineering Department,
Noble Group of Institutions, Junagadh-362001, Gujarat, India.

Dr. Vipul M. Vekariya
Associate Professor, Computer Engineering Department
Noble Group of Institutions, Junagadh-362001, Gujarat, India.

Abstract

Twitter is the most common social media and micro-blogging service, is a very popular method for expressing opinions. Twitter sentiment analysis is a challenging issue in this world. Sentiment Analysis for English language expanded in the field of sentiment analysis research in the recent years and its progress is quite high. But for Indian regional languages are it is still emerging approach and progress is too slow. In this paper, we present a practical approach to Twitter Sentiment Analysis. This approach is adopted to analyze Gujarati tweets and classify them into two basic polarities (Positive and Negative). As Pre-processing techniques improves classification accuracy, we focus on feature extraction. The experimental data is taken from Twitter. . Support Vector Machine (SVM) algorithm gains very well performance in sentiment analysis. In this work, we focus on POS tagging as a feature extraction and SVM to classify inputs.

Keywords: Sentiment Analysis, Machine Learning, Support Vector Machine, Gujarat Language

I. INTRODUCTION

Twitter, one of the most common online social media and micro-blogging services, is a very popular method for expressing opinions and interacting with other people in

the online world. Tweets (Twitter posts) are well-suited sources of streaming data for opinion mining and sentiment polarity detection.[1]

To analyze them classification is important task. For Gujarat, Gujarati is widely used language for speaking and writing. So Sentiment Analysis of Gujarati Tweets is a very useful task to be achieved.

As we have seen that lots of research found for English and some Indian languages. But there is little work found in Gujarati Language as it is the only language that is used in speaking and writing by most of people in Gujarat. When classification is about Gujarati script, it becomes complex task because Guajarati script contains many adjectives (વિશેષણ) and adverbs (ક્રિયાવિશેષણ).For Example: ખરાબ,ગંદા,ક્યારેક,ખુબ,સારા.

Sentiment analysis can be performed at three different levels: document, sentence and aspect level. The document level sentiment analysis can be defined as classification of entire document as positive and document. The sentence level sentiment analysis is similar to subjective method. At this level opinions (Positive, Negative and Neutral) are determined from each sentence. The aspect leave sentiment analysis aims at identifying the target of the opinion.[2]

II. SENIMENT ANALYSIS

Sentiment analysis can be defined as classification of text into different polarities. Following stages are used in sentiment analysis.

Figure 1: Sentiment Analysis Stages

Data Collection:

In this stage, data for sentiment analysis is collected from various social networks. For example, tweets from twitter, facebook comments, news reviews, movie reviews, etc.

Text Preparation:

This phase is also known as data Pre-processing phase. In this stage, data collected from above stage is pre-processed. For that various techniques are used like tokenization, stop words removal, stemming etc. Text preparation improves accuracy of classification [3].

Sentiment Detection:

Another name of this stage is Feature Extraction. Feature extraction and identification is necessary to identify sentiment polarity. This method extracts features from data for sentiment analysis. Feature selection techniques can be either filter based, wrapper based, embedded or hybrid, and can be used to evaluate individual features, or subsets of features. Filter based techniques are fast and scale well, making them well suited for tweet sentiment classification, as their speed is well matched with the high volume of instances and large number of features associated with this domain. By comparison, filter-based subset evaluation, wrapper based techniques and hybrid techniques require significantly more computational resources to be of use in this domain. Embedded techniques are inherent to a specific classifier and are not applicable when working with multiple diverse classifiers [4].

Sentiment Classification:

The basic principles of text classification based on machine learning methods are: the computer system automatically estimates the correlation between each text and various types of categories, then assign the text to one of the categories. Classification is done on the basis of three classes i.e. positive, negative and neutral. Various techniques are available like SVM, NB, K-NN, NLP, MLP, etc [5].

Presentation of Output:

There are three indexes generally used in text classification: Recall, Precision and Accuracy.[5]

III. PREVIOUS APPROACHES:

For sentiment analysis of English language, lots of work has been carried out. There are so many techniques used for sentiment analysis like Machine learning, Lexicon Based Method, Neural Networks, etc.We have analyzed various approaches like Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Decision Tree, Multi-Layer Perceptron, etc. for English language sentiment analysis. Comparisons of these approaches given in below table.

Table: 1: Comparison of Different Classifiers

Classifier	Advantages	Disadvantages
Support Vector Machine	Simple, good generalization capability, Gives better accuracy	High Computation complexity, limitation is in speed and size,
k-Nearest Neighbor	Simple, generally applicable and easy to include new training data,	computational complexity for classification, less accuracy

Bayesian Classifier	Obtain good results in most cases	High computational requirements, requires exact knowledge of class prior probabilities and class conditional probabilities of features.
Decision Tree	Training and classification time fast	Tree structure is not global optimum
Multi-layer Perceptron	Classification is fast	Long training time

IV. PROPOSED SYSTEM:

The following figure represents proposed model for our research work. There are various stages in proposed model which described below.

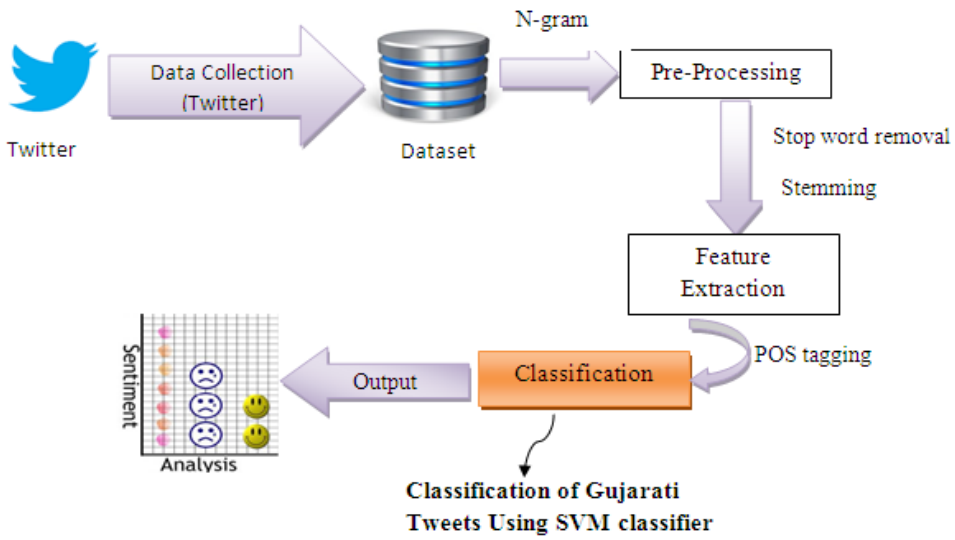

Fig.2 Proposed Model

Proposed Algorithm:

1) Create Data set:

Collect data from twitter and use these tweets for further steps. These tweets can be from any particular event like ચલણી નોટ, પાટીદાર આંદોલન, etc.

2) Divide tweets into words using N-gram algorithm:

Divide tweets into N-grams as like we have a sentence આ ફિલ્મ સારુ છે. Then N-gram algorithm gives output like this આ, ફિલ્મ, સારુ, છે.

3) Perform Pre-Processing:

Stop word removal removes Stop Words like આ, છે, જ્યારે, પણ by using common comparisons.

Perform stemming by using regular expression. Example of stemming is નોટના, નોટનું, નોટની, can be reduced to નોટ.

4) Feature Extraction using POS tagging:

For POS tagging, extract keywords Adjectives, Adverbs and Noun from remaining data.

5) Support Vector Machine to classify the inputs:

For classification using Support Vector Machine following functions are used:

$$\vec{w} * \vec{x_i} \geq 0, for y_i = +1(Positive)\ \vec{w} * \vec{x_i} < 0, for y_i = -1(Negative)$$

Where W is weight given in data dictionary

X is data vector.

V. RESULT ANALYSIS:

As we have seen that feature extraction improves classification accuracy, we have used POS tagging to extract features and SVM to classify input data. For this experiment we have used 40 tweets as a sample dataset. We have performed these different operations on our dataset. Fig.3 shows result of Support Vector Machine algorithm and Fig.4 gives number of classified positive and negative tweets. As a result we could achieve 92% accuracy.

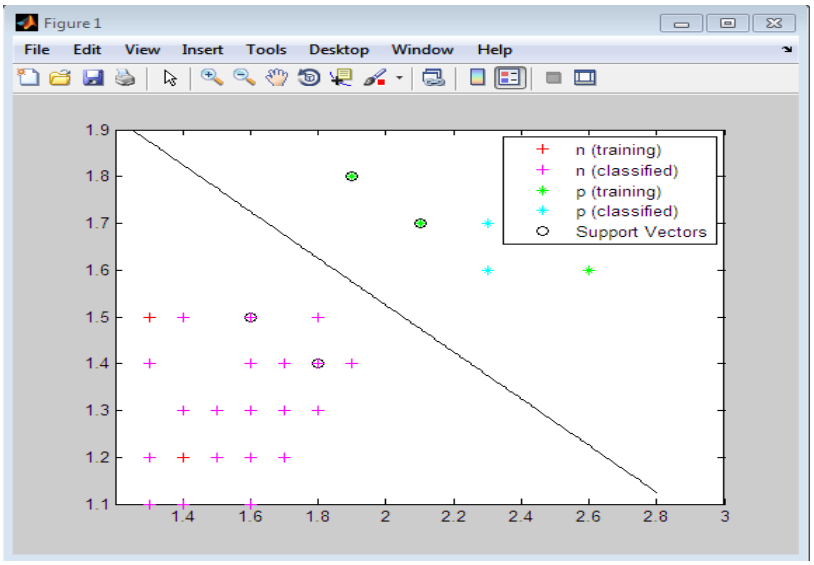

Fig.3 Results of SVM

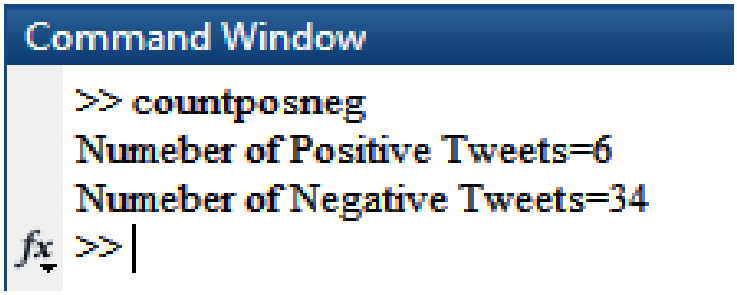

Fig.4 Counting number of positive and negative tweets

VII. CONCLUSION

This work includes classification of tweets into two basic polarities i.e. positive and negative. From literature survey, I found that SVM and POS tagging gives higher accuracy. So I will use POS tagging as a feature extractor and SVM as a classifier. The accuracy is the parameter that is used to measure the performance of the system. Results of this approach give accuracy 92%. This work can be extended to combination of any other feature extraction and machine learning approach.

REFERENCES:

[1] Peiman Barnaghi, John G. Breslin, Parsa Ghaffari, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment, 2016 IEEE Second International Conference on Big Data Computing Service and Applications, ISSN No.978-1-5090-2251-9/16 © 2016 IEEE.

[2] M.Trupthi, Suresh Pabboju, G.Narasimha, "Improved Feature Extraction and Classification -Sentiment Analysis",(IEEE) International Conference on Advances in Human Machine Interaction (HMI - 2016),March 03-05, 2016. ISSN No.978-1-4673-8810-8/16©2016 IEEE

[3] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Universite de Paris-sud, Laboratoire LIMSI-CNRS, Batiment 508, France.

[4] Joseph D. Prusa, Taghi M. Khoshgoftaar, Amri Napolitano, "Using Feature Selection in Combination with Ensemble Learning Techniques to Improve Tweet Sentiment Classification Performance", IEEE 27th International Conference on Tools with Artificial Intelligence, 2015.

[5] Wenying ZHENG, Qiang YE, "Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm", IEEE Third International Symposium on Intelligent Information Technology Application, 2009.

[6] Yakshi Sharma, Veenu Mangaty and Mandeep Kaurz, "A Practical Approach to Sentiment Analysis of Hindi Tweets", 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015), IEEE 2015.

[7] Deepika Kumawat, Vinesh Jain, "POS Tagging Approaches: A Comparison" (IJCA), Volume 118 – No. 6, May 2015

[8] Fang Luo, Cheng Li,Zehui Cao, "Affective-feature-based Sentiment Analysis using SVM Classifier", Proceeding of the 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design, ISSN No.978-1-5090-1915-1-/16 © 2016 IEEE.