Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques

Monisha Kanakaraj[1] and Ram Mohana Reddy Guddeti [2]

Dept. of Information Technology, National Institute of Technology Karnataka,
Surathkal, Mangalore 575025 India
kmonisha90@gmail.com[1]
profgrmreddy@nitk.ac.in [2]

***Abstract*—Mining opinions and analyzing sentiments from social network data help in various fields such as even prediction, analyzing overall mood of public on a particular social issue and so on. This paper involves analyzing the mood of the society on a particular news from Twitter posts. The key idea of the paper is to increase the accuracy of classification by including Natural Language Processing Techniques (NLP) especially semantics and Word Sense Disambiguation. The mined text information is subjected to Ensemble classification to analyze the sentiment. Ensemble classification involves combining the effect of various independent classifiers on a particular classification problem. Experiments conducted demonstrate that ensemble classifier outperforms traditional machine learning classifiers by 3-5%.**

***Index Terms*—Sentiment Analysis, Ensemble Classifier, Social Network Analysis, NLP Techniques**

I. Introduction

The Social Networking Sites (SNS) contain text, images, and other multimedia data. However, the short text messages posted by the users help in analyzing the attitude of the speaker on a particular topic thereby identifying the public opinion on the topic. To understand the view or opinion hidden in the text NLP is needed. With NLP the text data can be analyzed to mine the sentiment. Various Machine Learning algorithms such as Support Vector Machines (SVM), MaxEntropy, Naive Bayes are widely used to solve the classification problems. Ensemble methods in machine learning, combines the effect of multiple machine learning algorithms on the given problem set to obtain a better predictive power than its constituent algorithms in solitude. The Ensemble class of algorithms have many flavors depending on how the training datasets are chosen and on how the results are combined. The Averaging methods like Bagging, Forests of randomized trees provide the average of the base estimators as the ensemble output. While boosting methods such as Ada Boost, Gradient Tree Boosting give results which are the sequential effect of the base estimators. In this paper, we have analyzed the performance of Decision Tree, Random Forest, Extremely Randomized Trees and Decision Tree regression with Ada Boost Classifiers on Twitter sentiment analysis. In literature there are two main approaches for mining sentiment from the SNS: Dictionary Based (DB) and Machine Learning Based (MLB) as shown in Table I. In general all the DB techniques refer to a pre-built dictionary for classifying the sentiment. The potential limitation with DB systems is that, the strength of classification depends on the reference dictionary used. Also, most of these systems use Bag-of-Words concept which lacks domain/context based semantics. On the other hand MLB systems achieve domain based sentiment classification due to the presence of domain specific training data and the learning efficiency of classification algorithms. Class imbalance problem and linguistic variations in text can be overcome by the bootstrapping ensemble framework [6]. Coletta et al. [7] demonstrated the performance of SVM combined cluster ensemble classification on Twitter data.

TABLE I
EXISTING METHODOLOGIES

Type	Ref	Pros	Limitations
DB	[1]	Considers Lexicons & Emoticons	Lacks Domain Context
DB	[2]	Compares Online vs Social reviews	Lacks aspect rich data
DB	[3]	Bag of Objective Words	Dictionary dependency
MLB	[5]	NN combined SVM	Only on stock market data
MLB	[4]	Naive Bayes classifier	Baseline methodology

II. Proposed System

The proposed system gathers data from Twitter and does NLP techniques to frame the feature vectors. Then ensemble methods are applied on the training data to form the training model. Post training, the extracted feature vectors are classified based on the training model and the results are presented. The high level design of the proposed system is depicted in Figure 1. The system consists of the four main modules: Data Gathering module, Data Processing module, Training and Classification module and Classification output.

A. Data Gathering Module

The tweets are fetched using the Twitter API v1.1. The API v1.1 provides a more sophisticated programming interface through which the search results can be obtained in Object (called Tweet Object) format. This object format helps in extracting specific tweet attributes say user name, geographic location, re-tweet count etc. Then preprocessing of the gathered data is done to extract features.

IEEE ICSC 2015, February 7-9, 2015, Anaheim, California, USA
978-1-4799-7935-6/15/$31.00 ©2015 IEEE

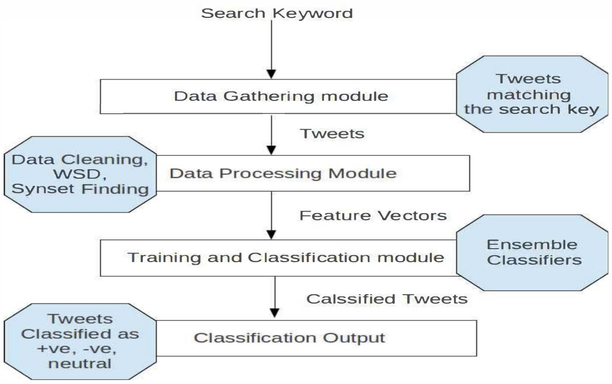
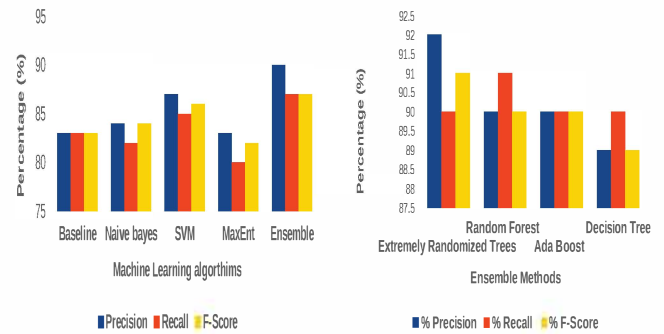


Fig. 1. Architecture of the Proposed System



(a) Comparison of ML algorithms (b) Comparison of Ensemble methods

Fig. 2. Results

B. Data Processing Module

1) Data Cleaning: Repeated letters and words removal are done in data cleaning module. For every keyword in the sentence, Word Sense tagging is done. URLs, hashed words, names are removed from the tweets. The cleaned tweets are now subjected to Parts Of Speech tagging.

2) Synset Finding: To capture semantic similarities among the tweets we use the synsets of WordNet. Synset contains the set of words that are semantically related to the word of interest. For every key word in tweet the synset of the word is retrieved from the WordNet database. This increases the accuracy of classification by covering all the semantically related data items. Before finding synset the original words in the tweets are stemmed up to the root word by user Stemmer. Stemming reduces the feature vector size while preserving the key terms.

3) Feature Vector Formation: The next step is to form the feature vector. The feature vector consists of key terms of the tweets along with the synset words.

The feature vectors thus obtained are subjected to classification by traditional classifiers and Ensemble classifiers and the results are presented to user stating the sentiment polarity of the public on the topic.

III. Experimental Results

Experiments were conducted to compare the performance of Ensemble method against other machine learning algorithms like SVM, Baseline, MaxEnt and Naive Bayes. Also experiments have been conducted to compare the results of various boosting and bagging ensemble methods. The performance parameters are : *Precision, Recall, F score*. We took an election dataset of 7086 classified tweets, two general tweet datasets with 1578627, 5513 classified tweets and movie review dataset of 25,000 tweets for training and testing. Figure 2(a) presents the comparative view of the performance of the traditional machine learning algorithms against the ensemble methods. It can be observed that, the precision of ensemble method outperforms the other solitary methods. In ensemble classification, each constituent learning method differ in the subset of the dataset used for training and the subset of the feature vector used for training and testing. These subsets are taken in random. It can be observed that Extremely Randomized Trees method performs better than other methods due to the high randomness in the way the splits are computed during feature vector subset selection. Unlike random forests where the the splitting threshold is chosen randomly, in extremely randomized tress most discriminative threshold is used as the splitting rule. Due to this, the variance of the ensemble on the whole decreases and hence the better performance as observed in Figure 2(b).

IV. Conclusion

The proposed system collects data from Twitter social network site and does NLP techniques to extract features from the tweets. Word Sense Disambiguation and WordNet sysnsets are used to increase the accuracy of prediction. Then various Ensemble methods of classification are applied to classify the data as Positive, Negative and Neutral. It has been observed that the ensemble method outperforms the traditional classification methods. Of the ensemble methods Extremely Randomized Trees classification performs better than others.

References

[1] P. T. Ngoc and M. Yoo, "The lexicon-based sentiment analysis for fan page ranking in facebook," in *Information Networking (ICOIN), 2014 International Conference on.* IEEE, 2014, pp. 444–448.

[2] A. Minanovic, H. Gabelica, and Z. Krstic, "Big data and sentiment analysis using knime: Online reviews vs. social media," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on.* IEEE, 2014, pp. 1464–1468.

[3] C. Hung and H.-K. Lin, "Using objective words in sentiwordnet to improve sentiment classification for word of mouth," *IEEE Intelligent Systems*, p. 1, 2013.

[4] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of facebook statuses using naive bayes classifier for language learning," in *Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on.* IEEE, 2013, pp. 1–6.

[5] A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis," in *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on.* IEEE, 2013, pp. 440–444.

[6] A. Hassan, A. Abbasi, and D. Zeng, "Twitter sentiment analysis: A bootstrap ensemble framework," in *Social Computing (SocialCom), 2013 International Conference on.* IEEE, 2013, pp. 357–364.

[7] L. F. Coletta, N. F. F. d. Sommaggio Silva, E. R. Hruschka, and E. R. Hruschka, "Combining classification and clustering for tweet sentiment analysis," in *Intelligent Systems, 2014 Brazilian Conference on.* IEEE, 2014, pp. 210–215.