Twitter Sentiment Analysis Using Hybrid Approach

Thakare Ketan Lalji[1], Sachin N. Deshmukh[2]

*[1, 2] Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad*

---***---

Abstract - *Strategic part of information gathering is to focus on how other people think. There are so many opinion resources that are rapidly growing and popular in world such as online review sites and personal blogs. In this paper we focus on Microblogging site Twitter. Twitter is a platform that allow user to express his opinion on variety of entities such as product, services, organization, people etc. We perform the sentiment analysis on those opinion tweets using Text Mining approaches such as Lexicon based approach and Machine Learning Approach. It performs Sentiment Analysis in two levels, first by searching the polarity words from the pool of words that are already predefined in lexicon dictionary and in Second level training the machine learning algorithm by polarities given by first level of lexicon*

Key Words: Sentiment Analysis, Social Media, Twitter, Lexicon Dictionary, Machine Learning Classifiers, SVM.

1.INTRODUCTION

Social Media sites such as Twitter, Facebook, Blogs are become important platform where user can share their valuable opinion on certain topics. With this kind of platforms various opportunities and challenges arise to actively use various techniques to extract and understand the opinion of others. Sentiment Analysis of twitter data has multiple usages like review of customer towards movie, product, services and application. Sentiment Analysis of tweets includes the classification of tweets as Positive, Negative or Neutral.

Lexicon Based Sentiment Analysis associate with the presence of certain word in document. Lexicon contains different features including the part of speech tagging of word, their sentiment values, subjectivity of word etc. The Sentiment Analysis of tweets are annotate using this features provided by these lexicons. Using that we can obtain polarity of whole tweet by averaging the sentiment values of words.

The Machine Learning based Sentiment Analysis technique requires creating a model by training the classifier with labeled examples. This means that first we require to gather a dataset with positive, negative and neutral classes, extract the features/words from that dataset & then train the algorithm based on the examples.

In this paper we use both these approaches Lexicon Based Approach and Machine Learning Approach. We show the result of sentiment analysis by combining these two approaches. Usually Lexicon based approach perform entity level sentiment analysis and it gives high precision but low recall. To improve the performance measurements such as Recall, F-Score, Accuracy. Machine learning algorithm is train using the polarity given by lexicon based approach. Our hypothesis is that the accuracy given by such approach is get increase with increase in size of training data.

The remainder of this paper is organized as follows. In section 2, we discuss the literature survey related to this paper. In section 3, we discuss the methodologies for twitter sentiment detection. In section 4, we show the results of experiment we done on different twitter datasets. In section 5, we discuss conclusion and future work.

2. RELATED WORK

The aim of our project is to classify the twitter messages as positive, negative or neutral. For this we use two approaches: Lexicon based approach and Machine Learning approach.

Lexicon based approach includes performing the sentiment analysis at document and sentence level by searching polarity of word from predefined word list.[1,2,3,4] determine polarity of sentence using predefined dictionary. Examples of such a Lexicon dictionaries are MPQA [5] and SentiWordNet 3.0 [6]

Machine Learning approach includes the three algorithms 1) Maximum Entropy, 2) Support Vector Machine (SVM), 3) Naïve Bayes (NB). This includes the training the classification algorithm [7].

In this paper we use combination of both approaches, it also called Hybrid approach. Normally combinations of both are used for subjectivity classification and then apply it to the learning algorithm [8]. Similar approach used in [9], which classify sentence in only two classes positive & negative, no neutral class it creates problem. [10] Uses same approach with different features. In [13] lexicon and machine learning approaches are combine. But they use different sentiment analysis methods. These are different than our approach; we first preprocess the data to remove unwanted data from it. Then we perform polarity detection using Lexicon dictionary and then apply this result to the Learning algorithm.

3. THE PROPOSED TECHNIQUE

Following figure 1 shows the workflow for the approaches we use to improve the performance of sentiment analysis.

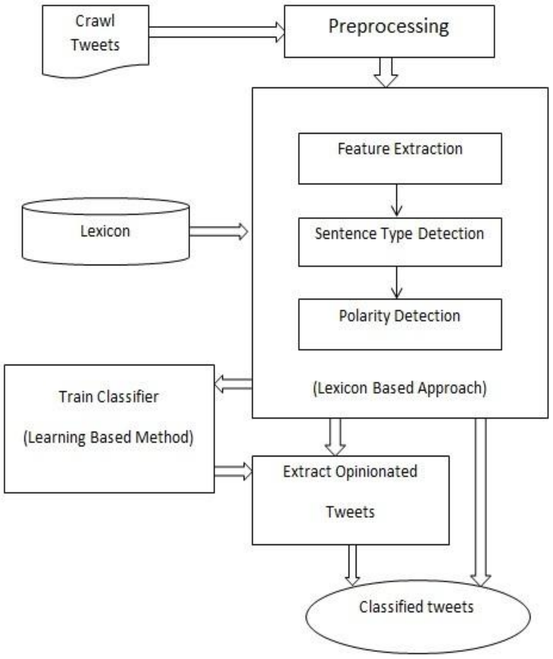

Fig -1: Workflow for sentiment Analysis

The Sentiment Analysis of tweets includes following steps in it.

3.1 Data Acquisition

Data acquisition is carried through the Twitter API. Twitter API allow user to interact to with its data i.e. tweets. User can download these tweets by creating twitter API. User request to API for the data and it returns data according to the query enter by user.

3.2 Pre-processing

The twitter data contain noisy data such as RT for Retweets, '#' hashtags for filtering tweets according to the topic, @usernames, external web links, and emoticons. The task of preprocessing removes all noisy data, so that data will be clean and it is easy to perform operations on clean data. We perform 1) Remove Duplicate tweets, 2) Remove Retweets, 3) Remove URL's, 4) Remove Unnecessary Space, 5) Remove twitter hashtags, 6) Remove Punctuation Marks, 7) Remove Numbers and, 8) Remove twitter username starts with @ symbol.

3.3 Feature Extraction

As tweeter contain much more unnecessary data, so we need to find data that contain opinion, which we use for sentiment analysis. So feature selection is way to find out this. Normally we find out the tweets that contain the Adjective, because presence of adjective in tweets indicate that the tweet contain the opinion about something in the world.

Next part in feature selection is to find out the subjective tweets. Subjective tweets are the tweets that contain the user emotion, view about something in the world. So it is necessary to find the subjective tweets, for that we need to classify tweets as Subjective and Objective tweets.

3.4 Polarity Detection

In this we just find the polarity of tweets by searching the occurrences of the word in the lexicon dictionary, and simple replace the word position with the polarity value shows by the lexicon dictionary. The Polarity of whole tweet is calculated by the aggregation of the word polarity present in that tweet.

3.4.1 Negation Handling

Negation Handling is major issue while sentiment analysis. Because many sentence contain the negation word that shifts the polarity of the sentence. Many classifiers remove the negation words by considering it as stop words. We had overcome this problem, when we find any negation term in sentence then we simply replace that negation term with punctuation symbol '!'. For that we had just made some changes in the lexicon simply add symbol '!' before each word in lexicon and just shifts the polarity of that words

Example:

Word 'Good' which has polarity is 1 i.e. Positive can be replace by '!Good' and set its polarity to -1 which indicate that it is a negation term

3.4 Sentiment Classifier

The polarity given by the lexicon dictionary for the each sentence can be considered as training data. These training data is given to Machine Learning Classifier to train the classifier. By training using this training data we calculate polarity of other data which can be passed as a testing data to the classifier. This implements the performance of the Sentiment Analysis.

4. EXPERIMENTS

4.1 Sentiment Classifier

Dataset: The initial task for sentiment analysis is the collection of dataset, we collect dataset using twitter API.

The query we fired while collecting data from API is 'car' i.e. the API gives all data (tweets) that contain the word car. For our experiments we collect 28000 tweets.

Preprocessing: Next task is to remove the noise form the collected data. Noise such as Duplicate tweets, Retweets, punctuations, numbers, HTML links etc.

4.2 Feature Extraction

Data we extracted contain unnecessary data. Hence to extract only those tweets that contain the some opinion, we perform feature selection. This includes extracting only those tweets that contain adjective. This can be done by using Tree Tagger Part of Speech Tagging (POS) technique [10]. And then we classify those tweets as Subjective and Objective tweets and consider only the Subjective as main feature for Sentiment Analysis. This can be performed using the MPQA Lexicon which contain the words with its subjectivity information i.e. whether word id Strong or Weak Positive [11]. After performing this feature selection we have 25000 tweets for further processing.

4.3 Polarity Detection

We can perform the polarity detection using the MPQA Lexicon [11], where we search for the occurrence of the each word of tweet in a lexicon dictionary, when find then replace that word with the polarity value given by the lexicon. When we find that word is not occur in the lexicon then we replace it with polarity value zero that indicate the Neutral polarity. Finally we aggregate all polarity values of words in tweets that the aggregate value indicates the polarity of the tweet. This tweets are consider as the training data, which are used to train classifier

4.3 Sentiment Classifier

In this section we train the classifier, so that it will assign the sentiment polarity to the newly opinionated tweets i.e. testing data. We use Support Vector Machine (SVM) as our leaning algorithm.

Training data

Training Data is data which can be labelled as positive, negative and neutral by the lexicon based method.

Classification Features:

Our basic features are Unigram, Bigram, and Trigram. We calculate the accuracy of polarity classification for the newly opinionated tweets using these classification features i.e. Unigram, Bigram and Trigram.

Test data

Testing data is newly opinionated tweet that are to be classified based on training given to the classifier using the data classified by the lexicon based method.

5. EMPIRICAL EVALUATIONS

For Evaluations we perform the sentiment analysis using the Support Vector Machine (SVM) learning method. The reason for using this algorithm is that it gives better result than learning algorithms like Naïve Bayes, Maximum Entropy [12].

Data Sets:

For the process of sentiment analysis, we divide training data into different parts, this is done to check the accuracy of sentiment classifier when the training data size increases. The aim is to just check the variations in the accuracy of sentiment classifier for same test data.

Evaluation Measures:

We use measure Accuracy to evaluate the sentiment classification performance. This measure can be check against classification features such as Unigram, Bigram and Trigram with different training size.

Evaluation Result:

Table 1 shows the accuracy for all three classification features, with variation in training data size.

Table -1: Accuracy Results

Training Data Size	Testing Data Size	Unigram	Bigram	Trigram
5000	1000	60.53	59.38	57.13
10000	1000	61.62	60.53	57.26
15000	1000	62.28	61.20	58.95
20000	1000	62.42	61.27	59.02
25000	1000	63.23	62.23	59.98

6. CONCLUSIONS

Sentiment Analysis for twitter data, when perform using Lexicon based approach it shows high precision but low recall so there is problem of performance. To increase that performance we combine both the approaches i.e. Lexicon Based Approach and Machine Learning Approach, this give better performance. We perform the empirical evaluation for the different sized training datasets. And these empirical results ensure that the proposed algorithm is highly effective and better for sentiment analysis of twitter messages.

REFERENCES

[1] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. 2011. Lexicon- based methods for sentiment analysis. Comput. Linguist. 37, (2): 267--307.

[2] Hu, M., & Liu, B. 2004. Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04). ACM, New York, NY, USA. pp. 168--177.

[3] Kim, S., & Hovy, E. 2004. Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING '04). Association for Computational Linguistics, Stroudsburg, PA, USA.

[4] Ding, X., Liu, B., & Yu, P.S. 2008. A holistic lexicon-based approach to opinion mining. In: Proceedings of the International Conference on Web Search and Web Data Mining (WSDM '08). ACM, New York, NY, USA. pp. 231-240.

[5] Pang, B., Lee, L., and Vaithyanathan, S. (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, EMNLP ’02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

[6] Multi Perspective Question Answering (MPQA). Online Lexicon “http://www.cs.pitt.edu/mpqa/subj_lexicon.html”.

[7] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani. “SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis s and Opinion Mining”. In Proceedings of international conference on Language Resources and Evaluation (LREC), 2010.

[8] Wiebe, J. and Rilo_, E. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. CICLing 2005.

[9] Tan, S., Wang, Y. and Cheng, X. 2008. combing Learn-based and Lexicon-based Techniques for Sentiment Detection without Using Labeled Examples. SIGIR 2008.

[10] www.cis.uni-muenschen.de/~schmid-tools/TreeTagger/

[11] Multi Perspective Question Answering (MPQA) Online Lexicon <http://www.cs.pitt.edu/mpqa/subj_lexicon.html

[12] Go, A., Bhayani, & R., Huang, L. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[13] Lei Zhang , Riddhiman Ghosh, Mohamed Dekhil,, Meichun Hsu, Bing Liu 2011. “Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis”. HPL Laboratories