International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

Social Media Sentiment Analysis: A Review

Princy Sharma[1], Vibhakar Mansotra[2]

[1]M.Tech Student, Department of Computer Science and IT, University of Jammu, J&K, India
[2]Dean faculty of Mathematical Sciences, Department of Computer Science and IT, University of Jammu, J&K, India

---

Abstract - *Nowadays, Thinking of people plays a very important role in different areas. The way people interact with each other express their opinions or sentiments. Sentiment analysis is one of the fastest growing research area in computer science. Sentiment analysis is used to identify users emotions towards some topics whether it is positive or negative. Social Media is the biggest platform for data sharing. The shared data is collected through social networking sites. Twitter is one of the most trending social networking site as people express their views on different topics. This paper presents an overview of different levels of sentiment analysis, twitter dataset and also, perform experiment on Twitter dataset.*

***Key Words*: Sentiment analysis, twitter dataset, Machine Learning.**

I. INTRODUCTION

The history of the Internet has changed the way people opinions on their perceptions. Now a days it is all done through different blog posts, product review websites and online discussion forums, etc [3]. People usually depend on user-generated content on any product to a great extent when it comes to perform any desired action. When they want to purchase a product through online, they will first look up its reviews in that particular product website through online sources, before making up a decision. Some analysis is to be done on all these reviews so that the final outcome says whether the product is good to buy or not. There are different sentiment analysis techniques that are available with many applications for different domains, like in education sector to get a feedback for teaching from students. Knowledgebase and Machine learning techniques are two techniques that are mainly used for sentiment analysis. In the case of Knowledge base approach this requires a large database with predefined emotions and an effective and efficient knowledge representation for identifying sentiments. In the case of Machine learning approach doesn't require any predefined set of emotions, this makes use of a training set in order to develop a sentiment classifier which classifies sentiments from the tweets and so machine learning approach is rather simpler than knowledgebase approach. There are different machine learning techniques that are used to classify data i.e., they are naïve Bayes classifier, support vector machine, decision tree, random forest, neural networks etc. There are also various Natural Language processing approaches are there one is the Lexicon based approach which is a rule based approach which depend upon two methods corpus based and dictionary based. Twitter is one of the most important social networking sites of today's generation. People express their views opinions related to various topic on twitter. Sentiment analysis is used to analyse the emotions of the various texts by differentiating tweets into positive and negative.

II. LEVELS OF SENTIMENT ANALYSIS

2.1. Document Level Sentiment Analysis

The Document Level Sentiment analysis is performed for whole document [4]. Single Document is a basic unit of information. In this type sentiment analysis, a single topic is considered for a review. But in case of forums or blogs, comparative sentences may appear and customers may compare one product with the other that has similar characteristics and hence document level analysis is not desirable in social media text which is in the form of blogs.

2.2 Sentence Level Sentiment Analysis

The Sentence level sentiment analysis is related to find sentiment form different sentences whether the sentence expressed is positive, negative or neutral sentiment. The Sentence level sentiment analysis is closely related to subjectivity classification. Here, the polarity of each sentence is calculated and then same document level classification methods are used for the sentence level classification problem. Then the objective and subjective sentences must be found out. The subjective sentences must contain opinion words which help in determining the sentiment about entity. After that the polarity classification is done into positive, negative and neutral classes [5].

2.3 Entity Level Sentiment Analysis

The Entity Level sentiment analysis performs finer-grained analysis. The goal is to find out the sentiment on entities or aspect of those entities. For example consider a statement "New Nokia cell phone has good picture quality but it has less battery backup." So the opinion on Nokia's camera and display quality is positive but the opinion on its cell phone battery backup is negative. The summery of opinions about entities can be created [6]. Comparative statements are part of the entity level sentiment analysis but deal with comparative sentiment analysis techniques.

2.4. Phrase Level Sentiment Analysis

In phrase level sentiment classification, the phrases that contain opinion words are found out and a phrase level classification is done. This is advantageous or may be disadvantageous. It is advantageous where the exact opinion about an entity can be correctly extracted [6]. But in other cases, where contextual polarity matters, result may not be accurate. So the negation of words can occur locally. In such cases, this type of sentiment analysis suffices [3].

2.5. Feature Level Sentiment Analysis

The features of product are considered as product attributes. Analysis of these features for identifying sentiment of the document is called as feature based sentiment analysis. In this approach of Sentiment Analysis positive, negative or neutral opinion is identified. The opinion is identified from the extracted features. It is the best analysis model among all other model.

III. TWITTER DATASET

Sentiment analysis is used to analyze opinions of users related to particular topic. Twitter has been most widely used social networking site where people express their opinions. The dataset from twitter helps us to know emotions of the people. The information present on twitter consist of only 280 characters in a single tweet[14]. For the extraction of dataset first, we create an account on twitter developer account and we send a request to twitter and requested them for API. API consists of keys and tokens. Twitter provide keys and tokens to the requester. The keys and tokens help the user to extract the tweets from the user. For the extraction of tweets we can use various languages like python and R. The collected tweets were consist of 15 variables (SNo., text, favourited, favourite count, replyToSN, replyToUID text, id, replyTOUID, statusSource, screenName, retweetCount, isRetweet, retweeted, longitude, latitude). The collected tweets were in JSON format. We can apply sentiment analysis on twitter dataset and evaluate positive and negative tweets by using various techniques.

IV. LITERATURE REVIEW

Umadevi in their paper titled, **"Sentiment Analysis Using Weka",** perform sentiment analysis of "SMS Spam collection Data Set" which consists of 5574 SMSs of positive and negative category by using Weka tool. Machine learning Algorithms used are Support vector machine and Decission tree and compare the performance of both the algorithms. It was concluded that Support Vector Machine is better than Decission Tree [10].

Aaquib Multani and Atul Agrawal in their paper titled **"Sentiment Analysis for Understanding Students' Learning Experiences: A Survey Paper",** provided a workflow for analyzing social data for educational purpose. Additionally, they proposed an approach for predicting sentiments of user specifies and classifying them in to 'negative' or 'positive'. They had implemented proposed method on JAVA environments using "Twitter" dataset [11].

Monalisa Ghosh and Goutam Sanyal in their paper titled, **"Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis"** performed sentiment analysis and investigated the inability or incompetency of the widely used feature selection methods(IG, Chi-square, and Gini Index) with unigram and bigram feature set on four machine learning classifcation algorithms (MNB, SVM, KNN, and ME). The proposed methods are evaluated on the basis of three standard datasets, namely, IMDb movie review and electronics and kitchen product review dataset [12].

Mohammad Mohaiminul Islam and Naznin Sultana in their paper titled, **"Comparative Study on Machine Learning Algorithms for Sentiment Classification",** performed sentiment analysis of text review using multiple machine learning algorithms. According to experimental results, Linear SVM approach is much better for sentiment classification. That assumption was made after a huge number of experiments by using different classifiers and combinations with two different review datasets [13].

Kavya Suppala and Narasinga Rao in their paper titled **"Sentiment Analysis Using Naïve Bayes Classifier"** performed sentiment analysis on twitter data by using a Naive Bayesian algorithm. By using model, they could measured the customers opinions and perceptions and enhanced them to any desired level depending on the data gathered from on line resources. In proposed work they used dataset from twitter and facebook [14].

Bac Le and Huy Nguyen in their paper titled **"Twitter Sentiment Analysis Using Machine Learning Techniques"** built a model to analyze the sentiment on Twitter using machine learning techniques by applying effective feature set and enhances the accuracy i.e., bigram,unigram and object-oriented features. The classification of tweets is done using 2 algorithms i.e., Naïve Bayes classifier and Support vector machines(SVM) whose accuracies are tested by calculating precision, recall and f-score and also shows same accuracy [15].

V. EXPERIMENT ON SENTIMENT ANALYSIS

Experiment on Sentiment analysis has been performed by using twiiter related tweets. Following are the steps followed for the sentiment analysis using WEKA tool:

5.1 Data Collection:

A Dataset has been collected from twitter. For the collection of tweets we had sent a request to twitter. After the approval of request API (keys and tokens) has been generated and we

were able to acess the Data. For this experiment, we had collected 20,000 tweets from the twitter. Those were general public tweets. Then we split the dataset into training and testing (70% for training and 30% for testing).

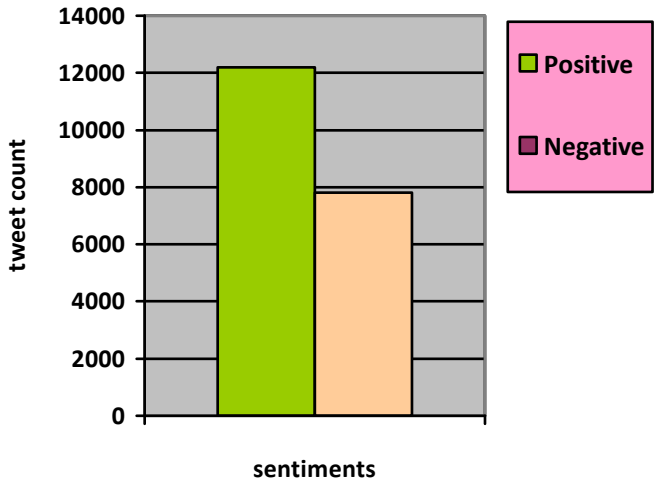
5.2 Data cleaning

The tweets extracted from the twitter were prone to inconsistency. Next step is data preprocessing, in which we clean the tweets by using various steps:

1. Remove URL.
2. Spelling Correction
3. Transform tweets into lower case.
4. Remove stop words
5. Stemming
6. Tokenization

5.3 Methodology

After data cleaning we used machine learning technique to find out the sentiment of tweets either they are positive or negative. For sentiment analysis we used a weka tool which classify twitter dataset and calculate the polarity of the tweets. For a given dataset polarity of the tweets is shown in the form of a bar graph.



Fig(1) Shows graph representation of positive and negative tweets

We used different machine learning algorithms to classify the tweets from twitter. In this experiment we considered 3 machine learning algorithms Naive bayes, Support Vector machine and Random Forest. Futher, we compared the performace of these algorithms. Below shown is the performance of different machine learning algorithm.

1. Random forest

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          51               89.4737 %
Incorrectly Classified Instances         6               10.5263 %
Kappa statistic                          0.7635
Mean absolute error                      0.2294
Root mean squared error                  0.3161
Relative absolute error                 50.1588 %
Root relative squared error             66.2057 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area
                 0.800    0.054      0.889     0.800     0.842     0.766     0.943     0.899
                 0.946    0.200      0.897     0.946     0.921     0.766     0.943     0.971
Weighted Avg.    0.895    0.149      0.894     0.895     0.893     0.766     0.943     0.946
```

Fig(2) random forest shows accuracy of 89.4 %

2. Decision Tree

```
Correctly Classified Instances          43               75.4386 %
Incorrectly Classified Instances        14               24.5614 %
Kappa statistic                          0.4729
Mean absolute error                      0.2996
Root mean squared error                  0.4014
Relative absolute error                 65.5028 %
Root relative squared error             84.0671 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area
                 0.700    0.216      0.636     0.700     0.667     0.474     0.818     0.722
                 0.784    0.300      0.829     0.784     0.806     0.474     0.818     0.884
Weighted Avg.    0.754    0.271      0.761     0.754     0.757     0.474     0.818     0.827
```

Fig(3) decision tree shows accuracy of 74.4 %

3. Naïve bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          53               92.9825 %
Incorrectly Classified Instances         4                7.0175 %
Kappa statistic                          0.8459
Mean absolute error                      0.0838
Root mean squared error                  0.2196
Relative absolute error                 18.3142 %
Root relative squared error             45.9409 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area
                 0.900    0.054      0.900     0.900     0.900     0.846     0.981     0.974
                 0.946    0.100      0.946     0.946     0.946     0.846     0.981     0.989
Weighted Avg.    0.930    0.084      0.930     0.930     0.930     0.846     0.981     0.984
```

Fig(4) Naïve bayes shows accuracy of 92.98 %

5.4 Results and discussion

We calculated sentiments of the tweets and the bar graphs shows the percentage of positive and negative tweets. After applying Random forest, decision tree and naïve bayes algorithms we concluded that naïve bayes shows higher accuracy than random forest and decision tree.

VI. CONCLUSION

Sentiment analysis is used to analyse public opinions either they are positive and negative. In this review paper various literature has been reviewed. This paper talks about different levels of sentiment analyses and also explained twitter dataset in details. The experiment has been done on Weka tool by using different machine learning techniques. The experiment shows among three proposed models naïve bayes algorithm has higher accuracy of 92.98%.

REFERENCES

[1] T. Singh et.al, "Current Trends in Text Mining for Social Media", International journal of Grid Distributed Computing, Vol. 10, No.6, pp. 11–28, 2017.

[2] L.Williams, et.al, "The role of idioms in sentiment analysis", Expert Systems with Applications, pp. 0957-4174, 2015.

[3] N. Munson, et.al, "Sentiment Analysis on the Social Networks Using Stream Algorithms", Journal of Data Analysis and Information Processing, pp. 60-66, 2014.

[4] Modha, G. S. Pandi, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 12, pp. 2277 128X, 2013.

[5] R. Varghese, J. M, "A Survey on Sentiment analysis and opinion mining", International Journal of Research in Engineering and Technology(IJRET), Vol. 2, pp. 2319-1163, 2013.

[6] S. B. Moralwar1, S.N. Deshmukh, "Different approaches of sentiment analysis", Computer Sciences and Engineering (ICSE), Vol. 3, pp. 2347-2693, 2015.

[7] N. Munson, et.al, "Sentiment Analysis on the Social Networks Using Stream Algorithms", Journal of Data Analysis and Information Processing, pp. 60-66, 2014.

[8] Rajaram, Ramasamy, and Appavu Balamurugan. "Suspicious E-mail detection via decision tree: A data mining approach." CIT. Journal of computing and information technology, Vol 15, no.2, 2007, pp. 161- 169.

[9] Ahmed, Ishtiaq, Donghai Guan, and Tae Choong Chung. "SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset." International Journal of Machine Learning & Computing, Vol 4, no.2 2014.

[10] Umadevi, "Sentiment Analysis Using Weka", International Journal of Engineering Trends and Technology (IJETT), Vol. 18, No.4, 2014.

[11] Aaquib Multani, Atul Agrawal, "Sentiment Analysis for Understanding Students' Learning Experiences: A Survey Paper", International Journal of Scientific Research & Engineering Trends, Vol 5, Issue 2, ISSN: 2395-566X, 2019.

[12] Monalisa Ghosh, Goutam Sanyal "Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis", Applied Computational Intelligence and Soft Computing, 2018.

[13] Mohammad Mohaiminul Islam, Naznin Sultana, "Comparative Study on Machine Learning Algorithms for Sentiment Classification", International Journal of Computer Applications, Vol. 182, No. 21, pp. 0975-8887, 2018.

[14] Kavya Suppala, Narasinga Rao, "Sentiment Analysis Using Naïve Bayes Classifier", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Vol. 8, Issue. 8 June, 2019.

[15] Le B., Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques", Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing, Vol 358, 2015.

[16] U. Ramakrishnan and R. Shankar, "Sentiment Analysis of Twitter Data: Based on User Behavior", International Journal of Applied Engineering Research, Vol. 10, pp. 16291-16301, 2015.

[17] S. Mathapati, S. H. Manjula and Venugopal, "Sentiment Analysis and Opinion Mining From Social Media: A Review", Global Journal of Computer Science and Technology, Vol. 16, pp. 77-90, 2016.

[18] L. Deng, J. Wiebe,"Recognizing opinion sources based on a new categorization of opinion types", International Journal of Computer Science and Information(IJCAI), pp.2775-2781, 2016.

[19] S. Shim and M. Pourhomayoun, "Predicting Movie Market Revenue Using Social Media Data", IEEE International Conference on Information Reuse and Integration, Vol. 04, no.1, 2017.

[20] Alsaeedi and M. Z. khan, "A Study on Sentiment Analysis Techniques Of Twitter Data", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 10, pp. 361-374, 2019.