Print ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Twitter sentiment analysis of game reviews using machine learning techniques

T.D.V. Kiran*, K. Gowtham Reddy, Jagadeesh Gopal
School of Information and Technology, VIT University, Vellore-632014, Tamil Nadu, India
***Corresponding author: E-Mail: tdvkiran11@gmail.com**

ABSTRACT

Sentiment analysis is basically analysing of the sentiments from the text. Sentiment analysis can be referred as opinion mining. Sentiment analysis finds and justifies the sentiment of the person with respect to a given source of content. The growth in micro-blogging activity on sites over the last few years has been increasing. Microblogging sites like twitter contain a large amount of data, which helps the companies to know what public is thinking of them. Sentiment analysis of this large data is very useful to express the opinion of the group of people. Twitter sentiment analysis is tricky as compared to broad sentiment analysis because of the slang words and misspellings and repeated characters. So it is very important to identify exact sentiment of each word. In our paper to obtain a highly accurate model of sentiment analysis of tweets with respect to latest reviews of Games which include both mobile and PC. With the help of feature selection and classifiers such as Support vector machine (SVM), Naïve Bayes and Maximum Entropy. By adding feature selection to the classifier we can select the relevant features. We will classify these tweets as positive, negative and neutral to give sentiment of each tweet and compare their results based on appropriate models.

KEY WORDS: Sentiment analysis, opinion mining, micro-blogging, Support vector machine (SVM), Naive Bayes, Maximum Entropy.

1. INTRODUCTION

With the increase in popularity of social media, a huge data is generated per second. It creates a lot of data with time. So the data generated by the social media or the microblogging sites are get wasted. But nowadays as the data analytics come into existence. The companies are making use of the data to generate the useful information from the data. Now in this paper, we are proposing a system for customers or gamers. This system reduces the time to analyze a large amount of data. Normally a user or gamer spends a lot of time in searching for the right product or the best product for them with respect to features or in a range of cost. This is the psychology of the customer. But this way of learning about the product takes a lot of time. And many of the reviews can be negative. But in social media people express their own feelings and they will not be any of the paid reviews types of things. So we are making use of the data from Twitter.

The people check the reviews or ratings of the games before playing that game on their PC or mobile. The quantity of information is unreasonable for a normal person to analyze with the help of naive technique.

Sentiment analysis is mainly concerned with the identification and classification of opinions or emotions of each tweet. Sentiment analysis is broadly classified into the two types first one is a feature or aspect-based sentiment analysis and the other is objectivity based sentiment analysis. The tweets related to game reviews come under the category of the feature based sentiment analysis. Objectivity based sentiment analysis does the exploration of the tweets which are related to the emotions like nice, bad, love, awesome etc.

In general, various symbolic techniques and machine learning techniques are used to analyze the sentiment from the twitter data. So in another way, we can say that a sentiment analysis is a system or model that takes the documents that analyzed the input, and generates a detailed document summarizing the opinions of the given input document. In the first step pre-processing is done. In the pre-processing we are removing the stop words, white spaces, repeating words, emoticons and #hash tags.

To correctly classify the tweets machine learning technique uses the training data. So, this technique does not require the database of words like used in knowledge-based approach and therefore, machine learning techniques are better and faster.

The several methods are used to extract the feature from the source text. Feature extraction is done in two phases: In the first phase extraction of data related to twitter is done i.e. twitters specific data is extracted. Now by doing this, the tweet is transformed into normal text. In the next phase, more features are extracted and added to the feature vector. Each tweet in the training data is associated with a class label. This training data is passed to different classifiers and classifiers are trained. Then test tweets are given to the model and classification is done with the help of these trained classifiers. So finally we get the tweets which are classified into the positive, negative and neutral.

Literature survey: In sentiment analysis we mainly categorize text data in sentiment polarity it may be either negative or positive. We have mainly three levels document level, sentence level and entity and aspect level. The document level concerns whether a document, as a whole, expresses the negative or positive sentiment, while the

sentence level deals with each sentence's sentiment categorization; The entity and aspect level then target on what exactly people like or dislike from their opinions.

2. RELEATED WORK

We will see some reviews on the previous topic based on our topic of research. Lina Zhou (2005), performed the investigation on game reviews using machine learning techniques. They used one of the technique as supervised vector machine to classify game reviews. A corpus is used to represent the data in documents and all classifiers are used to train the corpus. The proposed semantic orientation approach achieved 77% accuracy of game reviews. Supervised learning approach achieved 84.49% accuracy in three-fold cross validation. It achieved 66.27% accuracy on hold-out samples. It shows that semantic orientation approach is less accurate than supervised approach. It also concludes that supervised approach takes more time to achieve results.

Vishal A Kharde and Sonawane (2016) used few machine learning techniques to perform sentiment analysis on twitter data. They used three models maximum entropy support vector machine, and naïve Bayes. Maximum entropy classifier always tries to increase the entropy of the system by estimating the conditional distribution of the class label. It is represented as;

$$Pme(c|d,\lambda) = \exp(\sum \lambda f(c,d))/\sum \exp(\sum \lambda f(c,d))$$

It infers that maximum entropy has achieved an accuracy of 74.6% for unigram model. It stands between SVM and naïve Bayes of 77.73 % and 74.56 %. In this method, the relationship between features are taken irrespective of assumptions

Kamps, (2004) used the lexical database WordNet to determine the emotional content of a word along different dimensions. They developed a distance metric on WordNet and determined semantic polarity of adjectives. Dmitry (2010), proposed an approach to sentimental analysis based on the twitter user defined hashtag in tweets. For classification they have used the n-gram approach, punctuation, single words and pattern as different feature types and then combined in to a single feature vector for the classification. They have made use of K-nearest neighbour strategy to assign labels in each training and testing data set. They have concluded that SVM and Naïve Bayes are best techniques to classify the data and can be regarded as the baseline learning methods

Proposed method: In today's world due to continuous growth in the technology, there has been a rise in the quality of games and they have become more realistic to the current World. We are going to analyze the people's reaction based on one of the trending game. To analyze the data we are going to use twitter as the main information source. We use this data and apply few machine learning techniques on the data to extract the required knowledge.

The purpose of the paper is to determine the people's reaction accurately by comparing the results using few machine learning techniques.

Data Collection: The data is collected for dataset. The dataset regarding the attributes which suites the paper must be analyzed such that the entire result depends on the dataset and attributes present in it.

Data Pre-processing: Inconsistent data can lead to false conclusions .the data collected is analyzed for duplications, invalid values ,missing values ,corrupt or inaccurate values etc. this type of dirty data is then replaced, modified, deleted before further complications during mining .

Creation of feature vector: Features from tweets are extracted in two phases. In the first phase, features related to twitter are extracted. This is also called as extraction of twitter specific features. Then we have replaced the hashtags (#) with the exact same word by removing # sign, i.e. if the word is #PokemonGo replace it with PokemonGo. The Twitter specific features may not be present in all tweets. So we have further extracted the tweet to obtain more features. At this point, we have the tweet as the simple text. Then, using unigram approach, the whole tweet is represented by its keywords.

Data Classification: After creating a feature vector, classification is done using Naïve Bayes, Support Vector Machine, Max Entropy and the performance is compared.

Details of experimentation and modelling: The experimental details of the paper are given as follows:

Datasets: The full training dataset contains the 21,000 tweets and stored in the CSV file. Out of these, we are using 1200 tweets (600 positive tweets, 600 negative tweets, 600 neutral tweets) for training the classifiers. These datasets are collected from various sources and class labels are manually annotated whenever class labels are missing.

Pre-processing of Tweets: In the first step, the tweets are converted to lower case. So by doing this we can get words of each tweet in the same case (i.e. in lower case). Then in the next step, all the URLS are eliminated and replaced with normal text. Then we have replaced "@username" with generic word AT_USER. In the next step, we have removed the punctuation at the starting and ending of tweets and replace additional white spaces with single white space. After that #hashtag is removed with the exact same word, without the hash.

Modelling of Feature Vector: First, remove any stop words that are present in tweets. Then replace the character which is occurring more than twice in the particular word, with the two characters, i.e. trim the character which is repeated more than once. For example, replace "Superrrrrrrrr" with "Super" etc.

The examples of feature words extracted from sample tweets are shown below.

Table.1. Example Showing Tweets and Feature Words

Positive Tweets	Feature Words
Pokemon Go The game is exceptionally awesome	'awesome'

Negative Tweets	Feature Words
AT_USER disappointed. Playing game. It is a waste of time.	'disappointed', 'played', 'game', 'waste', 'time'

Neutral Tweets	Feature Words
Not the best game, but it's a time passing game	'not', 'best', 'game', 'but', 'time', 'pass'

Classification of Tweets:

Naïve Bayes Classifier: The classification using Naïve Bayesian is done as follows - First, all the tweets and labels are passed to the classifier. In the next step feature extraction is done. Now, both these extracted features and tweets are passed to the Naïve Bayesian classifier. Then train the classifier with this training data. Then the classifier dump file opened in write back mode and feature words are stored in it along with a classifier. After that the file is close.

Support Vector Machine: For SVM, we have basically used 3 labels that are 0, 1 and 2. Here the 0 represents positive, 1 represents negative and 2 as neutral. Each word in a tweet is represented as either 0 or 1. If it is feature word, then represent it with 1 otherwise 0. So we get a sequence of 0s and 1s. Now this feature vector and class labels are given to an SVM classifier to classify tweets as positive, Negative, Neutral.

Maximum Entropy: The Maximum Entropy classifier is a Probabilistic classifier which belongs to the class of exponential modes .It consists of taking probability distribution which maximizes information entropy, subject to the constraints of the information.

$$p_i = 1/n \text{ for all ' } i \text{ ' belongs to } \{1,\ldots\ldots,n\}$$

Retrieving tweets for a particular topic: By using the twitter account we have created the application for our paper and then the valid credentials are given to this JSON file.

We have defined config.json as below

{
"Consumer_key": "PROJECT PRODUCER KEY",
"Consumer_secret":"PROJECTPRODUCER SECRET",
"access_token":" PROJECT ACCESS TOKEN",
"access_token_secret":" PROJECT ACCESS TOKEN SECRET",
}

3. RESULTS

Since, we have used specific selected domain, there is no need of analyzing subjective and objective tweets separately. This show how the context or domain information affects sentiment analysis. As mentioned below SVM and Maximum Entropy have almost similar performance.

Maximum Entropy has higher accuracy than Naïve Bayes and SVM as well as better precision and recall. Maximum Entropy has accuracy of 90% while the other two are as follows SVM has accuracy 83.2%, and Naïve Bayes has 64.2%. This shows the quality of feature vector selected for this paper domain. The feature vector aids in better sentiment analysis despite of classifier selected. The accuracy of classification will increase as we increase the training data.

4. CONCLUSION

Thus we conclude that the machine learning technique is very efficient and easier than normal techniques. These techniques are easily applied to twitter sentiment analysis. Twitter sentiment analysis is difficult because it is very hard to identify emotional words form tweets and also due to the presence of the repeated characters, white spaces, slang words, misspellings etc. To handle these problems the feature vector is created. Before creating feature vector pre-processing is done on each and every tweet. Then features are extracted in two phases: First phase is the extraction of the twitter specific word. Then they are removed from the text. Now extracted feature vector is transformed into normal text.

After that, features are extracted from tweet which is normal text without any hash tags. And these extracted features are then added to form feature vector. There are different machine learning classifiers to classify the tweets. From our results, we have shown that Naïve Bayesian, Maximum Entropy and Support vector machine performs well and also provide higher accuracy. The results show that we get 64.2 % accuracy form Naïve Bayes, 83.2% accuracy form SVM classifier and 90 % accuracy from Maximum Entropy. So we can increase the accuracy of classification as we increase the training data. By this paper we can say that feature vector performs better for tweets related to Game reviews.

REFERENCES

Dmitry Davidov, Ari Rappoport, Enhanced Sentiment Learning Using Twitter Hashtags and Smileys, Coling, Beijing, 2010, 241-249.

Kamps J, Marx M, Mokken R.J and De Rijke M, Using wordnet to measure semantic orientations of adjectives, ELRA, 2004.

Lina Zhou, Pimwadee Chaovalit, Movie Review Mining, a Comparison between Supervised and Unsupervised Classification Approaches, Proceedings of the 38th Hawaii International Conference on system sciences, 2005.

Vishal A Kharde and Sonawane S.S, Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications, 139 (11), 2016, 5-15.