ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

Sentiment Analysis of movie Reviews using Twitter data

[1]Enimai V, [2]Gokulavani S, [3]Niveditha J P,[4]Varshini R, [5]Sheik Abdullah A
[1]-[4]UG Final Year, Department of Information Technology,
Thiagarajar college of engineering, Madurai.
[5]Assistant Professor, Department of Information Technology,
Thiagarajar College of Engineering, Madurai.

***Abstract:*-- Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker .It is used to identify the mood emotional tone of the speaker. Sentiment analysis is extremely useful in social media monitoring as it allows us to gain the idea of the public. The problem is hard to find the motive of the sentence. The data source is twitter using twitter API. The techniques used are term frequency(TF), inverse document frequency(IDF) and Support vector machine(SVM) which is used to seperate the positive, negative and neutral.**

***Keywords*: Sentiment analysis, term frequency, inverse term frequency, Support vector machine, twitter, twitter API.**

I. INTRODUCTION

Sentiment analysis is a method of analyzing peoples opinions, sentiments, evaluations, attitudes, and emotions from written language. Sentiment analysis systems are used in almost every domain because opinions are central to almost all human activities. They are key influencers of our behaviors .Due to the popularity of the social media such as blogs and social networking sites such as Facebook, Twitter etc the interest in sentiment analysis has increased to a higher extent. Sentiment analysis helps to find words that indicate sentiment and helps to understand the relationship between textual reviews and the consequences of those reviews. One such example being online movie reviews affect the box office collection. In this project, data mining techniques are applied on online movie reviews and predict the box office collection of the movie based on the reviews and analyze how much effect the reviews have on the box office collection. Box office collection for the next day is predicted based on online reviews of the present day. A prediction of high or low collection is also predicted.

It is one of the popular social networking site. It is a short text message of 140 characters. It has n number of users ie billions and billions of users. 500 million tweets are generated every day. But all users share their personal and social views about the movies. So the exact opinion about the movie can be judged using the tweets related the film.

In sentimental analysis we generally face several challenge. Normally to analyze data at a word level it is very easy. But to analyze it at a sentence level, it's a bit complicated. Also views are not presented in a similar fashion by all people. Word level sentiment analysis- It is used to find whether the sentence contain positive or negative meaning. It is used to find the polarity of the sentence i.e positive or negative. We chose twitter movie data because our aim is to find whether the movie is a success or failure.

Objectives

1. Use an optimize algorithm to classify the sentence.
2. Graphical representation of the sentiment score in the form of bar chart.
3. Consider a large amount of data sets for different movies.
4. Perform various levels of pre-processing.
5. Perform clustering to reduce the processing time.

The processing steps include,

1. Collection of data Sets.
2. Preprocessing
3. Finding TF (Term Frequency) and IDF (Inverse Document Frequency)
4. Clustering
5. SVM(Support Vector Machine) based Classification
6. Visualization of Results

The rest of the paper provides the following details: SectionII discusses the related work done in this domain. Section III explains the methodology in this paper in depth followed by the results and analysis obtained in Section IV and Section V gives the conclusion for the proposed work. Section VI describes the future works followed by references that we've used.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

II. RELATED WORK

Akshay et al, [1] proposed a Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques. by AkshayAmolik(2016), analyse 600 positive,600 negative and 600 neutral reviews are taken as training data set and 50 for above each as testing. Machine learning concepts and it resulted in 75 percent accuracy form SVM ,65 percent accuracy form Naive Bayesian classifier, can increase the accuracy of classification. Collection of larger datasets can be done has a part of future work. Umesh Rao Hodeghatta,[2]proposed Sentiment Analysis of Hollywood Movies on Twitter by (2015),Sentiment Analyzer tool - using python and natural language tool kit libraries by trying different supervised machine learning algorithms .Inclusion of more regions and the usage of other classification techniques, can be used as a future work

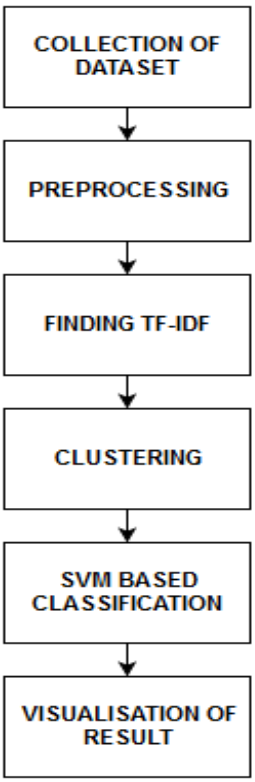
N PoongodiS,Radha[3] proposed Classification of user Opinions from tweets using Machine Learning Techniques by (2013) uses Natural Language Processing (NLP),Support Vector Machine(SVM), Naive bayes (NB) and Multilayer Perceptron (MLP) resulted in more accuracy. A twitter micro blog suffers from various linguistic and grammatical errors analysing those dataset is considered to be future work .

GeetikaGautam,Divakaryadav[4] proposed Sentiment Analysis of Twitter Data Using MachineLearning Approaches and Semantic Analysis by (2014) using Naive Bayesian, Maximumentropy and SVM along with the Semantic Orientation-based WordNet methods. The naive byes technique whichgives us a better result. Further the accuracy is againimproved when the semantic analysis WordNet is used.

Samad Hasan Basaria[5] proposed Opinion Mining of Movie Review using Hybrid Methodof Support Vector Machine and Particle Swarm Optimization uses Hybrid Method of Support Vector Machine and Particle Swarm Optimization resulted in accuracy of 52.31 percent. One of their future works is to experiment with different classifiers on our dataset.

Deepa Ananda[6] proposed Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering by using manual labeling (M), clustering(C) and review guided clustering (RC) resulted in analyzing of big data (tweets) only for text.The result of Text mining and data analysis would help in suggesting related pages based on different types of data. Further analysis can be done to images and all types of multimedia files.

III. METHODOLOGY



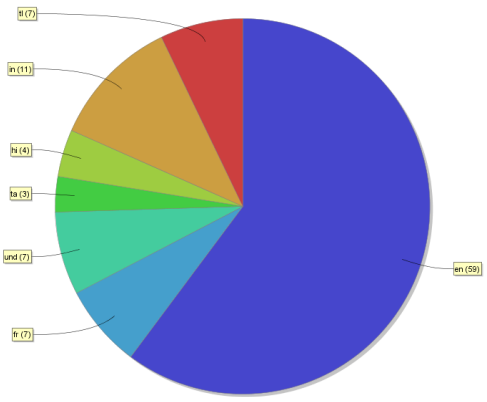
A .Collection of data set

The movie dataset collection is done with rapid miner using Twitter API . Creation of Twitter API is done by creating a user's twitter account. The Twitter API involves creation of an app called Movie Critic. Installation of Rapid Miner is done and a process is created . The linking of twitter API is done by setting up connection type as Twitter connection followed by giving access token and authorizing it. Query field is given as rapid miner and result type is recent and limit is given.Now a set of tweets is obtained and converted to .csv files.

Using Rapid Miner tool we have extracted tweets by English language .Below is the graph showing the distribution of Kabali movie by Language

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017



B. Preprocessing

Preprocessing is a first step needed to be done for efficient algorithm processing .Here, eliminating the noisy text from the retrieved data set or tweets. It has these steps:Removal of stop words like articles, prepositions is done mainly. Here, stop words means the most common words used in language like the, on ,which etc.

C. Finding TF-IDF

TF-IDF (Term Frequency and Inverse Document Frequency) is a technique used to categorize document. The algorithm categorize the movie reviews dataset. It generates the score of each word present in the Document.TF-IDF computes the weight which represent the importance of a term inside the document.

It increases proportionally to the number of times a word appears in the Document.TF-IDF is computed for each word in the Document

$$TF(t) = ND \div TD$$

Where ND means Number of times term t appear in a Document and TD means Total number of terms in the Document.

$$IDF(t) = \log[ND \div DF(t)]$$

Where ND means Total number of Documents and DF(t) means Number of Documents with term t in it.

D. Clustering

The above work for analysis of movie review using the frequency of word count as features for classification tends to provide result with lesser accuracy. So the review present in training should be analyzed in a better way.To overcome this problem clustering of review data based on TF-IDF measures was performed.

E. SVM based Classification

Support vector machine is a relatively new method of learning algorithm, that was initially brought to knowledge by Vapnik and co-workers (Boseretal, 1992;Vapnik, 1998) and was then extended by the other researchers. Their remarkable performance using the noisy data has made their works being used in a wide range of applications such as categorization of text and prediction. When the data is being put to use for classification, it separates the set of binary labeled training data with a hyper-plane that is maximally distant from them .

After we group the datasets under different clusters as per the number of clusters provided by the user by the usage of K-Means clustering, the obtained data is then used for classification sentimentally . By doing the clustering process, we get to divide the data in a better way and utilize the data in a more efficient manner. The final sentiment classification is done using the SVM classifier.For the two class problems SVM is generally put to use. During the training procedure to separate each class from the other we put to use the hyper plane formed. Using the space vector machine classification we feed the tweets and classify them as positive, negative and neutral ones. After classifying the tweets into their nature we then feed it in terms of a graph. This brings out the accuracyof the process in a accurate manner.

TABLE 1:Example showing Tweets and Featured word

Positive Tweet	Featured Word
With this being said, "Finding Dory" is also the third film by Disney to reach $1B. It is the third highest grossing film	'third film' ,'reach' , 'highest' , 'grossing'
Wishing team super luck for PremamGood reviews pouring in for Chay& the film Looks like they pulled it o…	'wishing','super','luck','good', 'pouring','like','pulled'

ISSN (Online) 2394-2320

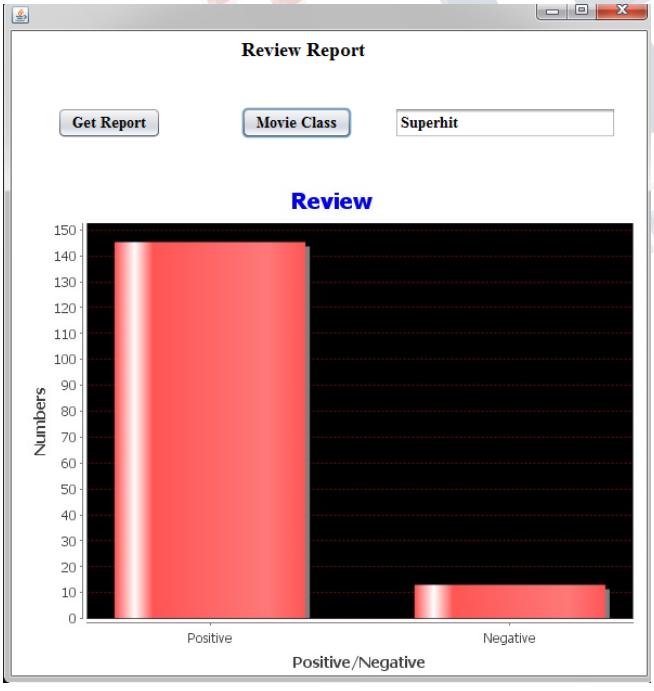
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

TABLE 2:Example showing Negative Tweet & Featured word

Negative Tweets	Featured Word
AT_USER disappointed. Watched a movie. It is a waste of time.	'disappointed','watched', 'movie','waste','time'

F. Visualization of results

The obtained results of tweets such as positive, negative and neutral ones are fed into any one of the data mining tools. Here, we've used rapid miner to analyze the tweets. As we insert the three categories of tweets into the tool, an output graph is produced that tells the user regarding the positive and negative response received from the audience. If the positive bar is high on the graph then the film hasreceived huge positive response from the people while if the negative bar is high on the graph then the film has received a negative impact from the society. As well, a resultstating if the film is a success or a failure is stated at the end of the process. After doing so the user gets to know if the film is worthy of its reviews or not is known to the user. Finally the accuracy measures for the tweets and respective reviews is obtained.



IV.RESULTS AND DISCUSSION

We have implemented the proposed model in Java (Netbeans IDE). The last outcome of the project is to find the number of positives and negatives and represent them in the form of Bar chart.At last it also gives the status of the Box-Office as good, moderate, excellent, hit and super hit.

The Evaluation metrics used for our project is Accuracy . It is a measure of predictive model which reflect the appropriate number of times the model is correct when applied to data. It is also defined as the common measure of Classification Performance. Accuracy can also be found as the proportion of the properly classified examples to the originally available total(examples). Since we use accuracy for the skew-data while using it one needs to be very careful.

The various formulae include,

$$precision = 100 \times (no.ofpositives \div (no.ofwords + no.ofpositives))$$

$$recall = 100 \times ((\ no.oftotalwords - no.ofpositives) \div no.ofpositives\ )$$

$$Fscore = (2 \times precision \times recall) \div (precision + recall)$$

$$Accuracy = no.ofpositives \div totalwords$$

V. CONCLUSION

In this paper we have done the implementation of our sentiment analysis of movie reviews by collecting tweets and processing them using an algorithm. The support vector machine is used to classify them and as a previous step the clustering is done. Now the analysis of a movie can be easily done by collection of opinion from a reliable social networking site such as twitter using Twitter API .The same application can be modified to collect and process review for a product in the similar way.

VI . FUTURE WORK

The application so for involves the class labeling process as ending along with visualization of results. The future work can be based on identifying bi-grams or negation inclusion for example:"I don't like this movie" is a sentence with a positive word "like" but a negation word "don't" is present before it ,used to determine the sentiment

by assigning a particular sentiment score to them and evaluating them accordingly.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classifi- cation using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002, pp. 79–86.

[2] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms," in Proceedings of the 2006 SIGCOMM workshop on Mining network data. ACM, 2006, pp. 281–286.

[3] A. Kyriakopoulou and T. Kalamboukis, "Text classification using clustering," in Proceedings of The 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Burlin, Germany, 2006, pp. 28–38.

[4] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002, pp. 129–136.

[5] "Imdb." [Online]. Available: http://imdb.com

[6] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in Information Communication and Embedded Systems (ICICES), 2013 International Conference on. IEEE, 2013, pp. 271–276.

[7] M. K. Jiawei Han, Data Mining:Concepts and Techniques. 500 Sansome Street, Suite 400, San Francisco, CA 94111: Diane Cerra, 2006.

[8] R. Yao and J. Chen, "Predicting movie sales revenue using online reviews." in GrC, 2013, pp. 396–40