Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques.

Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M.Venkatesan
School of Computer Science and Engineering,
VIT University, Vellore-632014, Tamilnadu, India
amolikakshay007@gmail.com
nike18tan@gmail.com
mahavirbhandari4@gmail.com
mvenkatesan@vit.ac.in

Abstract— Sentiment analysis is basically concerned with analysis of emotions and opinions from text. We can refer sentiment analysis as opinion mining. Sentiment analysis finds and justifies the sentiment of the person with respect to a given source of content. Social media contain huge amount of the sentiment data in the form of tweets, blogs, and updates on the status, posts, etc. Sentiment analysis of this largely generated data is very useful to express the opinion of the mass. Twitter sentiment analysis is tricky as compared to broad sentiment analysis because of the slang words and misspellings and repeated characters. We know that the maximum length of each tweet in Twitter is 140 characters. So it is very important to identify correct sentiment of each word. In our project we are proposing a highly accurate model of sentiment analysis of tweets with respect to latest reviews of upcoming Bollywood or Hollywood movies. With the help of feature vector and classifiers such as Support vector machine and Naïve Bayes, we are correctly classifying these tweets as positive, negative and neutral to give sentiment of each tweet.

Keyword- Feature Vector, Machine Learning, Twitter, Sentiment analysis, Unigram.

I. INTRODUCTION

With the increase in the popularity of social networking, micro-blogging and blogging websites, a huge quantity of data is generated. We know that the internet is the collection of networks. The age of the internet has changed the way people express their thoughts and feelings. The people are connecting with each other with the help of the internet through the blog post, online conversation forums, and many more. The people check the reviews or ratings of the movies before watching that movie in theatres. 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 in 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 movie 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 hate, miss, love 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 is 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 feature vector. Each tweet in the training data is associated with 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.

II. LITERATURE SURVEY

There are two techniques widely used to detect the sentiments from text. They are Symbolic techniques and Machine Learning techniques [3].

A. Sentiment analysis using Symbolic Techniques

A symbolic technique uses the availability of lexical resources. Turney [4] suggested an approach for sentiment analysis called 'bag of words'. In the mentioned approach, individual words are neglected and only collections of words are considered. He gathered word having adjectives or adverb for the polarity of review from a search engine Altavista.

A lexical database called WordNet [6] was used by Kamps et al [5] which determines an emotional matter in a word. WordNet carries synonyms and distance metric to find the orientation of adjectives.

To overcome obstacles in lexical substitution task, Baroni et al [7] developed a system supported by word space model formalism thereby representing local words.

EmotiNet conceptually represented the text that stored the structure of real events in a domain. This was introduced by Balahur et al [8].

B. Sentiment analysis using Machine Learning Techniques

Under this technique, there are two sets, namely a training set and a test set. Generally the dataset which is collected from different sources and whose behavior and output values are known to us falls into the category of training data sets. In contrast with this, the datasets whose values or behavior are unknown to us are called as test data sets. Here different classifiers are trained with training data and then unknown data or we can say a test data is given to this model to get desired results.

Machine Learning consists of various different classifiers such as Ensemble classifier, k-means, Artificial Neural Network etc. These are used to classify reviews [8].

Y.Mejova et al [1] in his research work proposed that we can use presence of each character, frequency of occurrences of each character, word which is considered as negation etc. as features for creating feature vector. He also shows that we can effectively use unigram and bigram approaches to make feature vector in Sentiment analysis.

Domingos et al [10] suggested that Naive Bayes works well for dependent features for certain problem. Zhen Niu et al [11] found a new model. This model is based on Bayesian algorithm. In this model, some efficient approaches are used for selecting feature, computation of weight and classification.

Barbosa et al [12] designed a 2 step analysis method which is an automatic sentiment analysis for classifying tweets. In the first step, tweets are classified into subjective and objective tweets. After that, in a second step, subjective tweets are classified as positive and negative tweets.

Celikyilmaz et al [13] developed one method as pronunciation based word clustering. This method normalizes noisy tweets. There are some words which have the same pronunciation but having different meanings. So, for eliminating this conflict, there is method mentioned above. In this mentioned method, words having same pronunciation are clustered and assigned common tokens.

Wu et al [14] in his paper recommended model, namely, the influence probability to analysis the sentiment tweets. In this, if @username is found in the tweet, it takes influencing action and helps to influencing probability.

By collecting automatic tweets, Pak et al [15] developed a method for sentiment analysis by creating twitter corpus. In his proposed work he shows that, while creating feature vector, we can use emoticons as a feature. He used a Naïve Bayesian classifier to do the sentiment analysis.

Some researches made to identify the public opinion about movies, news etc. from twitter tweets. V.M. Kiran et al [16] had taken the information from other publicly available databases like IMDB and Blippr.

III. PROPOSED METHOD

Various techniques have been used to do sentiment analysis of tweets. In our research we have used the method of feature vectors. The following Figure shows the entire proposed system architecture.

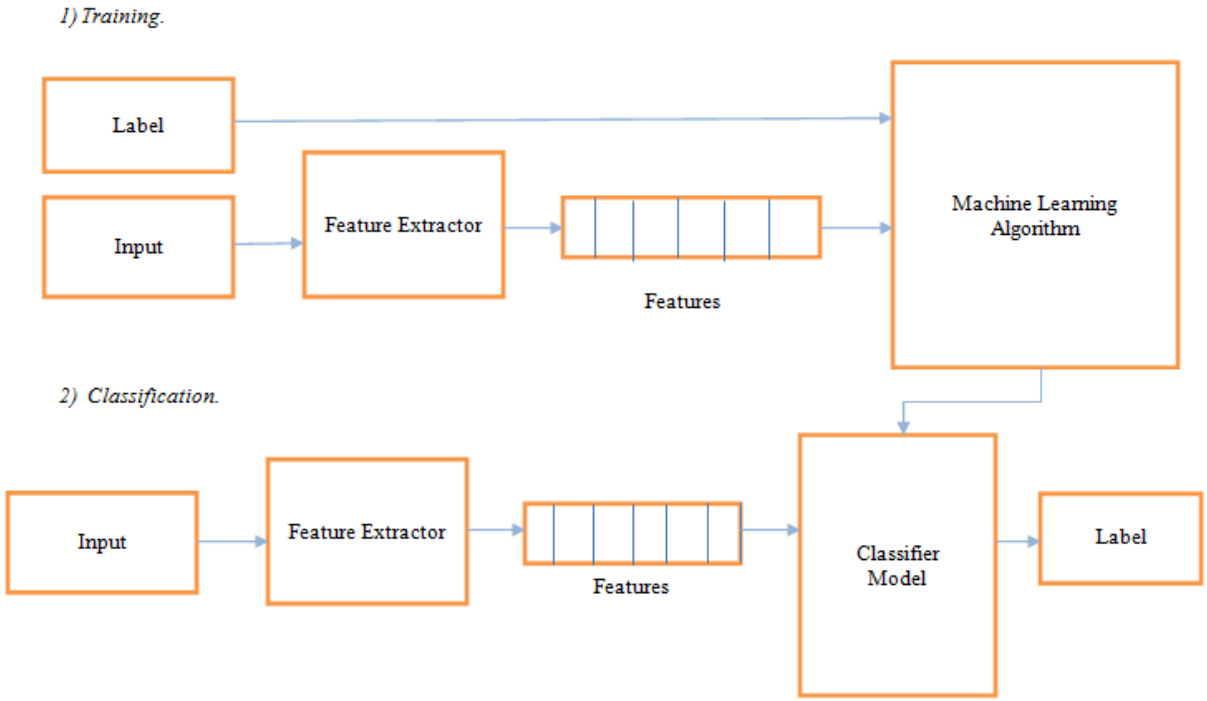

Fig. 1. System Architecture.

The proposed system contains various phases of development. A dataset is created using twitter posts of movie reviews. As we know that tweets contains slang words and misspelling. So we perform a sentence level sentiment analysis on tweets. This is done in three phases. In a first phase preprocessing is done. Then Feature vector is created using relevant features. Finally, using different classifiers, tweets are classified into positive, negative and neutral classes.

A. Creation of Dataset

- A dataset is created using twitter posts of movie reviews and related tweets about those movies.
- The below table shows dataset used for training the classifiers and also the tweets used for testing.

TABLE I. Statistics of the Dataset Used

Dataset	Positive	Negative	Neutral	Total
Training	600	600	600	1800
Testing	50	50	50	150

A dataset is created by taking 600 positive, 600 negative, 600 neutral tweets.

B. Preprocessing

The pre-processing is the major part of our project. In the pre-processing, first step is to convert the tweets into lower case. Next are to avoid the URL. Target name i.e. @username is replaced by using AT_USER and hashtags are removed. Next we have replaced the repeated character with the two occurrences and removed the white spaces.

C. 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 #BajrangiBhaijaan replace it with BajrangiBhaijaan. 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.

D. Sentiment Classification

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

1) Naive Bayes Classifier

The main advantage of Naive Bayes classifier is that it analyses each feature independently. So it makes the use of all the features in feature vector. The Probability of Naïve Bayesian classifier is given as,

$$P\left(\frac{z}{b_j}\right) = \prod_{i=1}^{m} P(z_i/b_j) \quad (1)$$

Where the feature vector is represented by z and b is the class label (i.e. positive, negative, neutral). Another reason of using Naïve Bayesian classifier is that it is simple to use and can be scalable. This classifier as compared to all other classifiers has high precision. But disadvantage of it is that the accuracy and recall is low. Naive Bayesian classifier is based on the famous Bayes theorem in mathematics.

2) Support Vector Machine (SVM)

With the help of large margins SVM does the classification. Each tweet is separated into single words. The descriptive function of SVM is given as below.

$$h(X) = z^x \emptyset(X) + c \quad (2)$$

Where the feature vector is represented by X. z suggests the vector of different weights. And non-linear mapping function is given by Ø and c is bias vector. Both z and c learn from the training data set automatically. SVM is one of the most highly accurate classifier in all other classifiers. In our project linear kernel has been used for classification. That is why it maintains the wide gap between the two classes. As compared to a Naïve Bayesian classifier, SVM has higher precision and recall.

IV. DETAILS OF EXPERIMENTATION AND MODELING

The experimental details of the project are given as follows

A. 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.

B. 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.

C. 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 "Smartttttttt" with "Smart" etc.

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

TABLE II. Example Showing Tweets and Feature Words

Positive Tweets	Feature Words
Bajrangi Bhaijann The film is exceptionally positive .Celebrate Humanity. Doesn't take any religion or country's side.	'positive', 'Humanity', 'religion', 'country's', 'side'

Negative Tweets	Feature Words
AT_USER disappointed. Watched a movie. It is a waste of time.	'disappointed',watched','movie','waste','time'
I miss my mom and dad. I hate this life.	'miss', 'hate'

Natural Tweets	Feature Words
Not a best movie, but one time you can watch it	'Not', 'best', 'movie', 'but', 'time', 'one', 'watch'
By by twitter, I am going to sleep now. Because tomorrow there is lots of work to do.	'sleep', 'now', 'tomorrow', 'lots', 'work', 'do'

D. Classification of Tweets

1) 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.

2) Support Vector Machine

For SVM, we have basically used 3 labels that are 0, 1 and 2. Here the 0 represents positive, 1 represent 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.

E. Retrieving tweets for a particular topic

By using the twitter account we have created the application for our project and then the valid credentials are given to this Json file.

We have defined config.json as below

```
{
  "Produser_key":  "PROJECT PRODUSER KEY",
  "Produser_secret":"PROJECTPRODUSER SECRET",
  "Control access_token":" PROJECT CONTROL ACCESS TOKEN",
 "Control access_token_secret":" PROJECT CONTROL ACCESS TOKEN SECRET",
}
```

V. RESULTS AND ANALYSIS

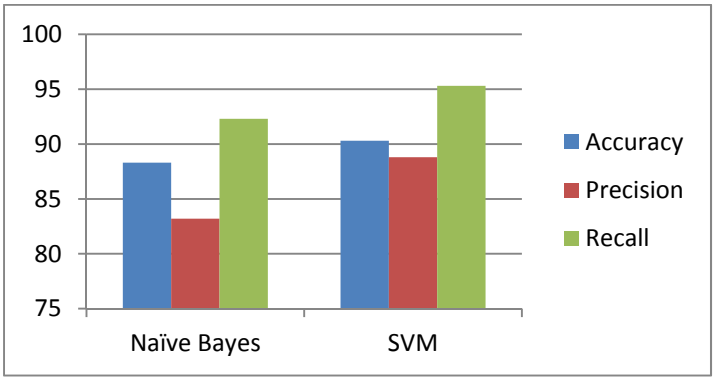

Fig. 2. Performance of classifiers in Twitter Sentiment Analysis.

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 shown in above graph SVM and Naïve Bayes have almost similar performance.

Naïve Bayes has better precision compared to SVM, but slightly lower accuracy and recall. SVM has higher accuracy than Naïve Bayes as well as better precision and recall. SVM has accuracy 75%, while that of Naïve Bayes has 65%. This shows the quality of feature vector selected for this project 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.

The performance of the system depends on training datasets and also content (i.e. Tweets) in these data sets. Thus, this is very simple and effective approach to analyze the sentiment form text. If we add other classifiers such as Maximum Entropy and Ensemble classifier we can easily analyze these results.

VI. CONCLUSION

Thus we conclude that the machine learning technique is very easier and efficient than symbolic techniques. These techniques are easily applied to twitter sentiment analysis. Twitter sentiment analysis is difficult because it is very tough to identify emotional words form tweets and also due to the presence of the repeated characters, slang words, white spaces, misspellings etc. To handle these problems the feature vector is created. Before creating feature vector pre-processing is done on each 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 or slang words. 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 and Support vector machine performs well and also provide higher accuracy. The results show that we get 75 % accuracy form SVM and 65% accuracy form Naïve Bayesian classifier. So we can increase the accuracy of classification as we increase the training data. By this project we can say that feature vector performs better for tweets related to Movie reviews.

ACKNOWLEDGMENT

The authors of this paper would like to thank Mr. Venkatesan M (Associate Professor, School of Computer Science and Engineering, VIT University, Vellore) for his support on this research work. Also, we gratefully acknowledge the contribution of VIT University to provide this wonderful opportunity and good facilities to carry out this research work.

REFERENCES

[1] Neethu M, S, Rajasree R,'Sentiment analysis in Twitter using Machine Learning Techniques', 4th ICCCNT , 2013.
[2] Y. Mejova, 'Sentiment analysis: An overview', ymejova/publications/CompsYelenaMejova, vol. 2010-02-03, 2009, 2009.
[3] E. Boiy, P. Hens, K. Deschacht and M. Moens, 'Automatic sentiment analysis in on-line text', 11th International Conference on Electronic Publishing, vol. 349360, 2007.
[4] P. Turney, 'Thumbs Up or Thumbs Down? Semantic orientation applied to unsupervised classification of reviews', 40th annual meeting on association for computational linguistics, vol. 417424, 2002.
[5] J. Kamps, M. Marx, R. Mokken and M. De Rijke, 'Using wordnet to measure semantic orientations of adjectives', 2004.
[6] C. Fellbaum, 'Wordnet: An electronic lexical database (language, speech, and communication)', 1998.
[7] D. Pucci, M. Baroni, F. Cutugno and A. Lenci, 'Unsupervised lexical substitution with a word space model', Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence, Citeseer, 2009.
[8] A. Balahur, J. Hermida and A. Montoyo, 'Building and Exploiting Emotinet, a knowledge base for emotion detection based on the appraisal theory model', Affective Computing, IEEE Transactions, vol. 3, 188101, 2012.
[9] G. Vinodhini and R. Chandrasekaran, 'Sentiment analysis and opinion mining: A survey', International Journal, vol. 2, 6, 2012.
[10] P. Domingos and M. Pazzani, 'On the optimality of the simple bayesian classifier under zero-one loss,',Machine Learning, vol. 29, 2-3, 103130, 1997.
[11] Z. Niu, Z. Yin and X. Kong, 'Sentiment classification for microblog by machine learning,', Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, vol. 286289, 2012.
[12] L. Barbosa and J. Feng, 'Robust Sentiment Detection on Twitter from Biased and Noisy data', 23rd International Conference on Computational Linguistics: Posters, vol. 3644, , 2010.
[13] A. Celikyilmaz, D. Hakkani-Tur and J. Feng, 'Probabilistic Model-Based Sentiment Analysis of Twitter Messages', Spoken Language Technology Workshop (SLT), 2010 IEEE, vol. 7984, 2010.
[14] Y. Wu and F. Ren, 'Learning sentimental influence in twitter', Future Computer Sciences and Application (ICFCSA), 2011 International Conference,IEEE, vol. 119122, , , 2011.
[15] A. Pak and P. Paroubek, 'Twitter as a Corpus for Sentiment Analysis and Opinion mining', Proceedings of LREC, vol. 2010, 2010.
[16] V. Peddinti and P. Chintalapoodi,V.M.Kiran, 'Domain adaptation in sentiment analysis of twitter', AnalyzingMicrotext Workshop, AAAI, 2011.

AUTHOR PROFILE

Akshay Ram Amolik received the Bachelor of Engineering degree in computer science and engineering in 2012 from Pune University, Pune, and Maharashtra, India. In 2013, he had received a Post Graduate Diploma degree in wireless and mobile computing from the Center for Development of Advanced Computing (C-DAC), Pune, Maharashtra, India. He is currently working towards the M.Tech degree at School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. He has a work experience of two years as ANDROID DEVELOPER in Reliance Industries, Mumbai, India. He has developed various android applications for Android phones, Android Tablets and Android TV. His research interest includes Big Data Analytics, Data Mining, Mobile Computing and Wireless Technologies, Computer Vision, Operating System, Semantic learning and Visual inference.

Niketan Janakraj Jivane received the Bachelor of Engineering degree in computer science and engineering in 2014 from Smt. Kashibai navale college of Engineering, Vadgaon (BK), Pune, Pune University, Pune, Maharashtra, India. He is currently working towards the M.Tech degree at School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. His research interest includes Big Data Analytics, Data Mining, Cloud Computing, Language Processing, Computer Networks, Semantic learning.

Mahavir Kundanmal Bhandari received the Bachelor of Engineering degree in Information Technology in 2014 from Solapur University, Solapur, Maharashtra, India. He is currently working towards the M.Tech degree at School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. His research interest includes Image Processing, Big Data Analytics, Data Mining, Computer Architecture and Computer Networking, Operating System.

Dr. M. Venkatesan is working as an Associate Professor in School of Computing Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu. He is working in the area of spatial data mining. His research area includes databases, data warehouse, big data analytics, data mining and applications of data mining in disaster management. He has worthy knowledge in data mining tools like Clementine, Rapid miner. He is the principal investigator in the ISRO funded project on data mining in a landslide. He completed his BE, MTech and PhD in the area of spatial data mining. .He published 25 papers in international journal and 30 papers in the reputed conferences.