# A Survey on Twitter Sentiment Analysis of Bollywood Movie Reviews

**Mr. N. S. Magar,**
M.Tech Final Year-Department of CSE-Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, (MH), India
Email - nitinmagar09@gmail.com

***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 280 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. With the help of feature vector and classifiers such as Naïve Bayes and Random forest we are correctly classifying these tweets as hit, flop and average to give sentiment of each tweet.*

***Key Words:*** *Feature Vector, Machine Learning, Twitter, Sentiment analysis, Unigram.*

## 1. 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.

## 2. LITERATURE REVIEW:

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

### *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] .which determines an emotional matter in a word. Word Net carries synonyms and distance metric to find the orientation of adjectives. To overcome obstacles in lexical substitution task, et al [7]

### *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] there 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.

## 3. PROPOSED WORK DESIGN :

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

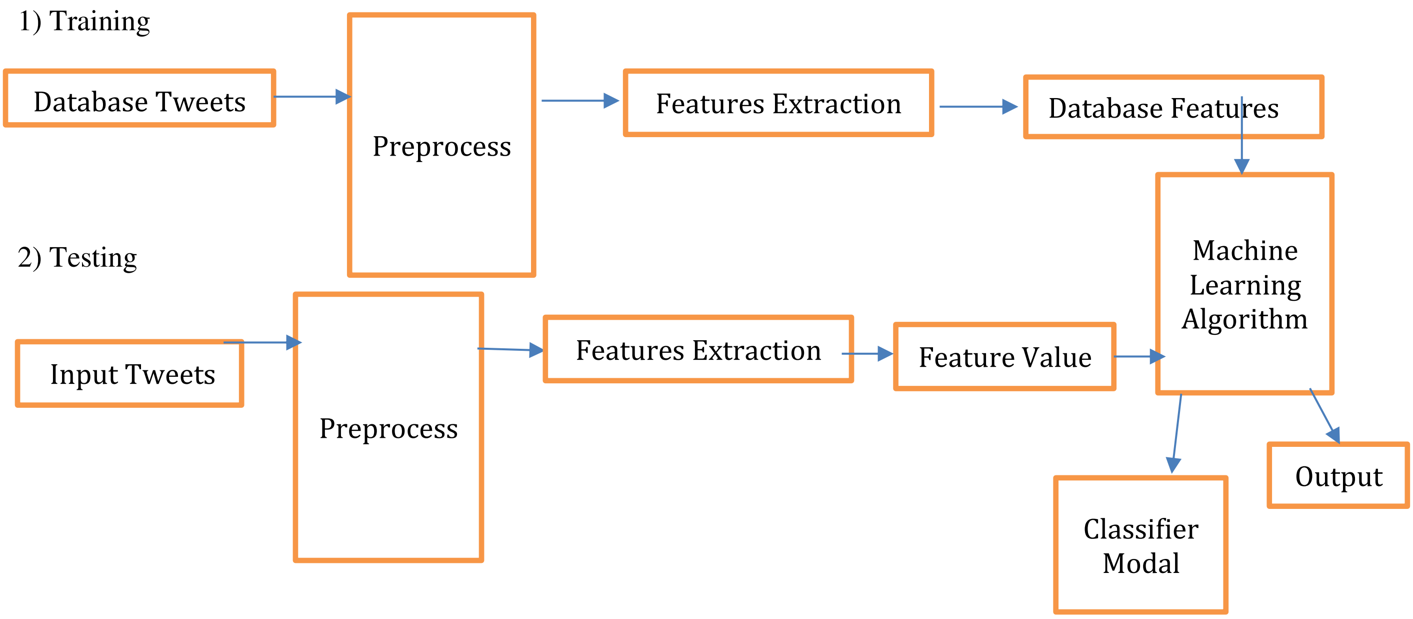

Fig1 : Propose work Design.

**Training**

The process of **training** an ML **model** involves providing an ML algorithm (that is, the learning algorithm) with **training** data to learn from. The term ML **model** refers to the **model** artifact that is created by the **training** process. ... For general information about ML **models** and ML algorithms, see Machine Learning Concepts

**Testing**

Here, once the model is obtained, you can predict using the model obtained on the training set. When given a data set for which you want to use Machine Learning, typically you would divide it randomly into 2 sets. One will be used for training, the other for testing. If you have a number of different algorithms and you wish to know which works best for this particular domain, train them all on the training set and see how they perform on the test set.

## 3. 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. In this paper we use machine learning technique and find out sentiment analysis of Bollywood movie reviews.

## 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,.
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, 2010.
16. V. Peddinti and P. Chintalapoodi,V.M.Kiran, 'Domain adaptation in sentiment analysis of twitter',