Suman Rani *et al*, International Journal of Computer Science and Mobile Applications,
Vol.5 Issue. 10, October- 2017, pg. 83-91 **ISSN: 2321-8363**
Impact Factor: 4.123

SENTIMENT ANALYSIS OF TWEETS USING SUPPORT VECTOR MACHINE

Suman Rani[1], Jaswinder Singh[2]
[1]M.Tech. Scholar, GJU & ST, Hissar, suman.sigar10@gmail.com
[2]Assistant Professor, GJU & ST, Hissar, jaswinder_singh_2k@rediffmail.com

Abstract: Sentiment analysis is the method of automatic mining of the sentiments or opinions of a text unit. In today's world, people express their feelings, opinions on twitter about certain things i.e. event, topic or person. The major focus of sentiment analysis is to categorize the opinions of the people by analyzing the views posted by them on social media. The proposed work deals with mining sentiments or emotions of tweets about Indian politicians using Support Vector Machine. Unigram and TF-IDF are used as feature extractors and the performance of the proposed approach is measured in terms of accuracy, precision, recall, and f-measure.
Keywords: Sentiment Analysis, Opinion Mining, SVM, Twitter.

1. Introduction

In today's era, the impact of the social networking media such as Facebook, Google Plus, YouTube, Blogs, and Twitter is increasing rapidly day by day. Millions of people are connected with each other on social networking sites and express their sentiments and opinion through tweets, and comments [1]. This motivates the automatic mining and classification of views, emotions, opinions, and feeling of people on social networking websites. Sentiment analysis is the process of analyzing the data in order to extract sentiment or opinions. It is also known as subjectivity analysis, opinion mining and sentiment classification. An example of sentiment analysis is online shopping. Online sites are full of product reviews. Before buying a product, a customer read reviews about that product. With the help of sentiment analysis, customers can find opinions of other people, whether they are satisfied or not with the quality of the product. Sentiment analysis is a type of natural language processing task that tracks the views of people about a certain thing or topic and categorizes these views into two classes i.e. positive and negative. In **positive class,** positive opinion of the authors is reflected like "this is a good movie." In **negative class,** the negative opinion of the authors is reflected like "this place is ugly."

Sentiment analysis also referred as Opinion mining can be applied at different levels i.e. Document level, Sentence Level, and Feature level [2] [4]. At Document level, the overall sentiments expressed in the whole document are classified into positive or negative class about a particular object. It is considered that full document contains opinions on a single entity. Both supervised and unsupervised learning techniques can be applied at the document level. At the sentence level, the opinion presented in each sentence is classified into positive or negative class. It considers that each sentence has only one sentiment. At Feature level known as aspect level sentiment analysis at

first, the aspects of the entity are determined, and then the polarity of the object is determined whether it is positive or negative.

Sentiment analysis can be done either using supervised learning techniques or unsupervised learning techniques [3]. The **Supervised learning** also known as machine learning approach use the data set that is divided into training set and test set. Naïve Bayes, Maximum Entropy, and Support Vector Machine are various techniques that come under the domain of supervised techniques. In **Unsupervised learning,** there is no need of dividing the data set into training and test data set. Lexical Resources like WordNet and SentiWordNet are the most widely used techniques for unsupervised learning approach. Among these techniques, Support Vector Machine is proven to be an excellent approach for text categorization [13].

The main aim of this work is to apply Support Vector Machine for analysis of tweets about Indian Polyticians and categorizing the tweets into positive and negative categories. Feature extraction has also been employed with support vector machine to improve the performance of classifier.

2. Literature Survey

Researchers have applied various techniques and tools [2] [3] [4] [16] for sentiment analysis.

Bholne Savita D. and Deipali Gore [5] have used Foreground and Background Latent Dirchlet Allocation (FB-LDA) model and Reason Candidate and Background Latent Dirchlet Allocation (RCB-LDA) model for the sentiment analysis of Twitter data. FB-LDA model purifies the foreground tweets and removes the background tweets. RCB-LDA model extracts the most relevant tweets from purified foreground tweets and finds out the reason behind sentiment variation. Sentiment labeling is assigned with help of Twitter Sentiment and SentiStrength tools. Support Vector Machine algorithm is preferred for sentiment classification.

Rincy Jose and Varghese S Choorali [6] proposed a lexicon based approach for sentiment analysis of twitter tweets. Sentiment towards two politicians Arvind Kejriwal and Kiran Bedi is compared during Delhi election days. Sentiment classification is done by lexical resources SentiWordNet and WordNet. To increase accuracy, Word Sense Disambiguation and negation handling is also introduced

Payal B. Awachate and Vivek P. Kshirsagar [7] conducted an experiment of twitter sentiment analysis. Three different categories of feature selection schemes are used. Among these categories, the combination of n-grams and sentiment lexicon dictionary gives the highest performance. For sentiment classification, three machine learning models Naïve Bayes, Decision Tree, and Kernlab Support Vector Machine are applied. Out of these classifiers, the Kernlab SVM gives the highest accuracy as compared to others.

V. S. Pagolu *et al.* [8] predicted the stock market movement of Microsoft Company from the extracted twitter data. They evaluated the correlation between changes in stock prices of the company and public sentiments conveyed in tweets about that company. Feature extraction is employed using n-grams and word2vec methods. The Random Forest algorithm is used for sentiment classification of the tweets.

Geetika Gautam and Divakar Yadav [9] analyzed the customers' reviews based on the twitter data. A set of machine learning techniques i.e. Naïve Bayes, Maximum Entropy, SVM and semantic analysis are applied for sentiment classification. Extracted adjectives from the dataset are used as a feature vector. The performance of the classifier is measured in terms of accuracy, recall, and precision.

Anurag P. Jain and Mr. Vijay D. Katkar [10] presented a method to predict sentiments tendency of people towards political situation and issues using data mining classifiers (k- nearest neighbor, RandomForest, BayesNet and Naïve Bayesian). Two lack tweets are gathered about various political parties and leaders using twitter API v1.1. The accuracy of the single classifier is compared with an accuracy provided by an ensemble of classifiers.

Rincy Jose and Varghese S Choorail [11] and Monisha Kanakaraj and Ram Mohana Reddy Guddeti [12] implemented sentiment analysis of twitter data using ensemble classifier. The classifier assemble approach is a combination of various classifiers. The accuracy of sentiment classification is improved using ensemble classifier over individual classifier.

3. Proposed Work

In the proposed work, SVM classifier is applied to mine the opinion of people about Indian politicians. First, the data is collected in the form of tweets from twitter using twitter API then these data is preprocessed, after that the features are extracted using Unigram and TF-IDF feature extraction method. Then the data is classified using SVM classifier. The proposed work is divided into following phases.

A. Data Collection: It is the first step of the sentiment analysis. Blogs, movie reviews, product reviews, email, and social networking sites are rich sources of data. We collected data from twitter using twitter API. For accessing twitter data, we have to create an app using twitter account which provides credentials (consumer key, consumer secrets, access token and access token secret).

B. Preprocessing of Data: Preprocessing includes the following things.

- All URL (e.g. www.xyz.com), hash tags (e.g. #topic) and targets (@username) are removed.
- Uppercase letters are converted into lowercase.

- All the texts are broken down into tokens. This process is called tokenization. For example "this is an amazing phone" is broken into individual tokens such 'this', 'is', 'an', 'amazing' and 'phone'. On encountering a space, a token is identified.
- Stop words like articles, prepositions, conjunctions, and pronouns are removed. Stop words provide little or no information.

C. Feature Extraction: It is the most important tasks related to classification. It includes the removal of irrelevant words or terms that do not express any sentiment. Unigram (n=1) [14] and term frequency and inverse document frequency (TF-IDF) [15] is used for feature extraction. The unigram represents individual and distinct words. The TF-IDF assigns a score to each word. The term- frequency is computed by counting the number of times a given word or term appeared in given document and inverse document frequency is computed by dividing the total number of documents by number of documents that has a given term.

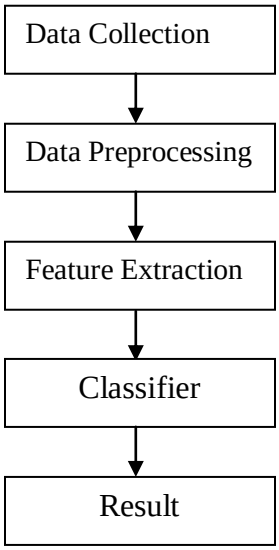

Figure 1*:* Steps of proposed work

D. SVM Classifier: SVM was introduced by Vapnik and others in 1992. It was originally designed for binary classification. It can also be extended to multi-class by combining multiples SVMs. It gives an excellent result for text categorization tasks such as sentiment analysis [13]. Here we consider binary SVM for simplicity. SVM performs classification by finding an optimal hyper-plane that separates two

classes. The optimal hyper-plane has maximum margin. The distance between nearest data point and hyper-plane is called as margin. The point that lies nearest to hyper-plane is called support vector.

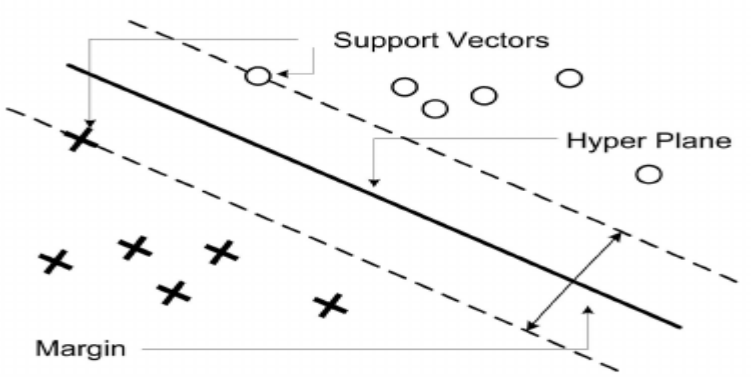

Figure 2: SVM Classification

Linear case: The dataset is represented as $D = \{(x_i, y_i | x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^n$ Where D is dataset of n rows that consisting of element (x_i, y_i). This dataset is a pair of x_i and y_i where x_i having a p dimensional feature vector and y_i having label of classes. Decision functin that separates two classes is defined as

$$f(x) = sign(w^T.x + b)$$

Where w is normal vector and b is the intercept of the hyperplane. w and b specifies the position of the hyperplane. The objective of hyperplane to search the optimal separating hyperplane. Mathimatically SVM problem can be formulated as

$$Minimize\ \frac{1}{2}w^T w$$

$$subjected\ to\ 1 - y_i(w^T.x_i + b) \leq 0\ for\ any\ i = 1, \ldots\ldots, n$$

The optimal hyperplane can be found by solving above linearly constrained quadratic optimization equation. On solving this equation, we get following solution:

$f(x) = sign(\sum_{i,j=1}^n \alpha_i\, y_i x_i x_j + b\ )$

Where α_i is is Lagrange multiplier.

Non-Linear case: For non-linear classification, SVM transforms the original data into a higher dimension. Then it seeks the linear optimal separating hyperplane or decision boundary in a a new dimension. The decision function is given as $f(x) = sign(\sum_{i,j=1}^n \alpha_i\, y_i \Phi(x_i).\Phi(x_j) + b\ )$. Where Φ is non-linear mapping function.

4. Experimental Setup and Result Analysis

The proposed algorithm is implemented in Python [17] with NLTK [18]. About 3745 tweets of people about three Indian politicians Narender Modi, Rahul Ghandhi and Arwind Kejriwal are collected. Each tweet is labelled as positive class or negative class. Two types of Support Vector Machine are applied for classifications of twitter data. Eighty percent data are used for training and twenty percent data are used for testing.Table 1 shows the types of support vector machines.

Technique	Library	Description
LinearSVC	Liblinear	linear SVM
NuSVC	libSVM	kernelized SVM

Table 1**: Support Vector Machine**

The performance of the proposed algorithm is analyzed using four parameters i.e. Accuracy, Recall, Precision, and F-measure.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - Measure = \frac{2 * Precison * Recall}{Precision + Recall}$$

Here, True positive (TP) defines the number of positive tweets that are correctly classified, as positive, whereas false positive (FP) is the number of negative tweets that are incorrectly classified as positive. True negative (TN) is the number of negative instances that are correctly classified as negative and false negative (FN) is the number of positive tuples that are incorrectly classified as negative tweets.

4.1 Results

Figure 3 shows positive and negative sentiments or views of public towards three Indian politicians. Different leaders have different sentiment results according to their working procedure. Among these leaders, Narender Modi is more successful politician.

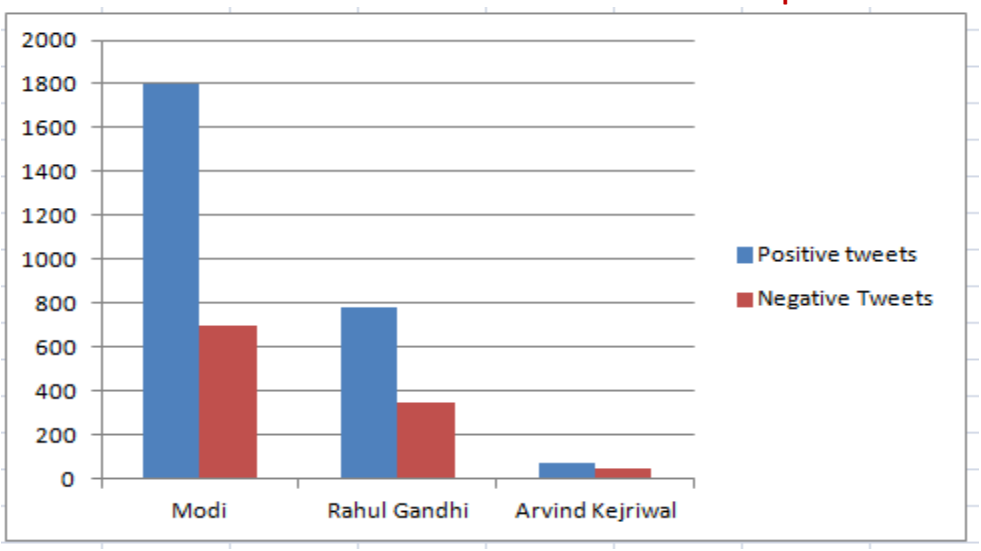

Figure 3: Sentiment Analysis of opinions of people about Narender Modi, Rahul Gandhi, and Arvind Kejriwal

Two types of SVM are applied for classification of tweets. Table 2 shows the experimental results of linear SVM and Kernel SVM. Figure 4 shows the Performance Comparison between Linear SVM and Kernel SVM.

Technique	Accuracy	Precision	Recall	F-measure
Kernel SVM	90	81	90	85
Linear SVM	93	90	90	90

Table 2: Experiment result of Linear SVM and Kernel SVM

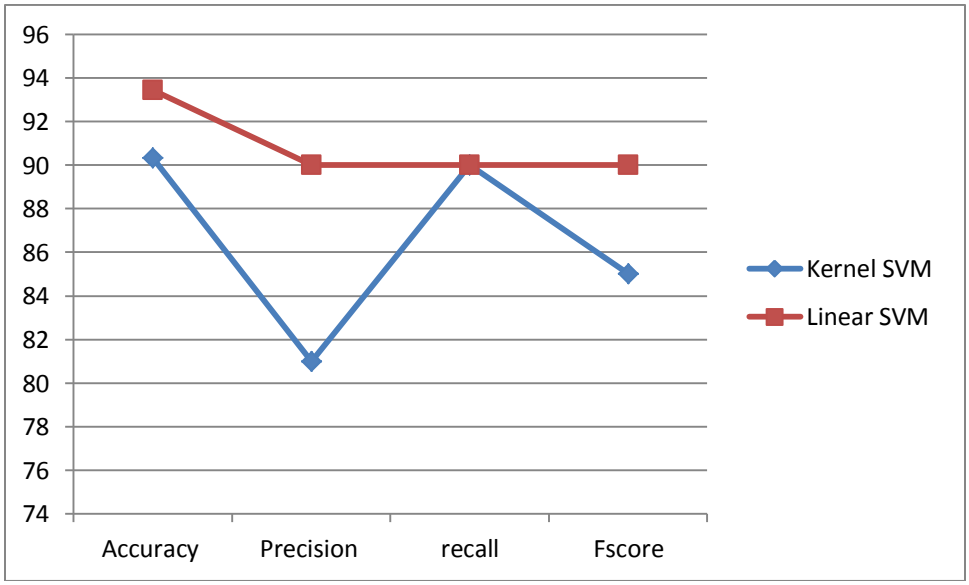

Figure 4*:* Performance Comparison between Linear SVM and Kernel SVM

It has been analyzed that the Linear SVM gives more accuracy as compared to Kernel SVM. The proposed approach is compared to other sentiment analysis technique that is based on the unsupervised learning approach [6].The unsupervised learning approach gives 78.6% accuracy whereas our approach gives above 80% accuracy.

5. Conclusion

Sentiment analysis is the process of categorizing a person's opinion and emotions expressed in the form of text. In today scenario, social networking sites are full of people opinions about product reviews, movie reviews, politics or a particular topic, etc. In this paper, linear SVM and Kernel SVM are applied for classification of the sentiments of about three Indian politicians. The tweets of people are collected from twitter using twitter API. In addition to it, feature selection is also applied here. The performance of both SVM is analyzed on various measures, i.e, accuracy, precision, recall, and f-measure. It is analyzed that linear SVM provides better performance than kernel SVM on all the measures. The proposed approach is also compared with another unsupervised sentiment analysis approach. It is concluded that the proposed approach performs better than the unsupervised sentiment analysis approach. In future, a hybrid technique of machine learning and lexicon technique can be used to sentiment analysis.

References

[1] A. U. R. Khan, M. Khan, and M. B. Khan, "Naïve Multi-label Classification of YouTube Comments Using Comparative Opinion Mining," *Procedia Computer Sci.*, vol. 82, pp. 57–64, 2016.

[2] D. Alessia, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, tools, and applications for sentiment analysis implementation," *Int. J. Computer Appl.*, vol. 125, no. 3, 2015.

[3] A. Kaur and N. Duhan, "A survey on sentiment analysis and opinion mining," *Int. J. Innov. Adv. Comput. Sci.*, vol. 4, pp. 107–116, 2015.

[4] Asmita Dhokrat, Sunil Khillare and C. Namrata Machender, "Review on Techniques and Tools used for Opinion Mining," *International Journal of Computer Applications Technology and Research*, vol. 4, no. 5, pp. 419-424, 2015.

[5] D. Bholane Savita and D. Gore, "Sentiment Analysis on Twitter Data Using Support Vector Machine," *International Journal of Computer* Science Trends and Technology, vol. 4, no. 3, 2016

[6] Rincy Jose and Varghese S Chooralil, *"Prediction of Election Result* by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation," *International Conference on Control, Communication & Computing India (ICCC),* pp. 638-641, 2015.

[7] Payal B. Awachate, Vivek P. Kshirsagar, "Improved Twitter Sentiment Analysis Using N Gram Selection and Combinations," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 5, 2016

[8] V. S. Pagolu, K. N. R. Challa, G. Panda, and B. Majhi, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements," *International conference on Signal Processing, Communication, Power and Embedded System (SCOPES),* pp. 1345-1349, 2016.

[9] G. Gautam and D. Yadav, Sentiment Analysis of Twitter Data Using Machine Learning Approaches And Semantic Analysis," *2014 Seventh International Conference on Contemporary Computing (IC3),* pp. 437–442, 2014.

[10] A. P. Jain and V. D. Katkar, "Sentiments Analysis of Twitter Data Using Data Mining," *2015 International Conference on Information Processing (ICIP)*, pp. 807–810, 2015.

[11] R. Jose and V. S. Chooralil, "Prediction of Election Result by Enhanced Sentiment Analysis On Twitter Data Using Classifier Ensemble Approach," *International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pp. 64–67, 2016.

[12] M. Kanakaraj and R. M. R. Guddeti, "NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers," *2015 3rd International Conference on Signal Processing, Communication And Networking (ICSCN)*, pp. 1–5, 2015.

[13] Z. Liu, X. Lv, K. Liu, and S. Shi, "Study on SVM Compared with the other Text Classification Methods," *2010 Second International Conference on Education Technology and Computer Science*, pp. 219–222, 2010.

[14] C. Tillmann and F. Xia, "A Phrase-Based Unigram Model for Statistical Machine Translation," *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pp. 106–108 2003.

[15] K. Ghag and K. Shah, "SentiTFIDF–Sentiment Classification using Relative Term Frequency Inverse Document Frequency," *Int. Journal Adv. Computer. Sci. Appl.*, vol. 5, no. 2, pp. 36–43, 2014.

[16] Suman and Jaswinder Singh," Sentiment Analysis: A Survey," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol 5, no. 8, 2017.

[17] http://www.python.org/downloads/

[18] http://pypi.python.org/pypi/nltk

