# SENTIMENT ANALYSIS USING TWITTER DATA

**Aditya Pratap Singh[1], Ashutosh Singh[2], Gopal Ji Pandey[3]**

*[1,2,3]Student, Dept. of Computer Science Engineering, RKGIT, Ghaziabad, Uttar Pradesh, India*

---

**Abstract –** *Social networks are the key tools for collecting knowledge about people's thoughts and feelings on various subjects as they spend hours on social media every day and express their views. In this technical paper we demonstrate how sentimental analysis is applied and how to link to Twitter and conduct sentimental analysis queries. We run experiments on different queries from politics to humanity and show interesting results. We noticed that the neutral feeling for tweets is substantially high, showing clearly the weaknesses of the current works.*

## 1. INTRODUCTION

Opinion and sentiment mining is an important research area because due to the huge number of daily posts on social networks, extracting people's opinion is a challenging task. About 90 percent of today's data has been provided during the last two years and getting insight into this large scale data is not trivial [17, 18]. For example in industries, sentimental analysis has many applications for various domains to get feedback for goods from which companies can gain customer feedback and social media reviews. Opinion and sentimental mining were well studied in this article, and all the different approaches and study areas were discussed [10]. Some research has also been done on Facebook but in this paper we concentrate more on the emotional study of Twitter. For larger texts one solution could understand the text, summarize it, and give weight to it whether it is positive, negative, or neutral. Two fundamental approaches to extract text summarization are an extractive and abstractive method. In the extractive method, words and word query on different topics and show the polarity of tweets.

### 1.1 RELATED WORKS

There are two basic methodologies to detect sentiments from text. They are Symbolic techniques and Machine Learning techniques [2]. The next two sections deal with these techniques.

### 1.2 SYMBOLIC TECHNIQUES

Much of the research in unsupervised sentiment classification using symbolic techniques makes use of available lexical resources. Turney [3] used bag-of-words approach for sentiment analysis. Relationships between the individual words are not regarded in that approach, and a text is interpreted as a pure word set. To determine the overall sentiment, sentiments of every word is determined and those values are combined with some aggregation functions. He found a review's polarity based on the average semantic orientation of tuples extracted from the analysis where tuples are adjective or adverbial phrases. Using search engine Altavista he noticed the semantic orientation of tuples. Kamps et al. [4] used the lexical database WordNet [5] to determine the emotional content of a word along different dimensions. They built a distance metric on WordNet and calculated the adjectives' semantic orientation. WordNet database is composed of terms linked by reciprocal relationships. Baroni et al. [6] developed a system using word-space model formalism that overcomes the difficulty in the lexical substitution task. It, along with its overall distribution, represents the local context of a word. Balahur et al.

[7] introduced EmotiNet, a conceptual representation of text that stores the structure and the semantics of real events for a specific domain. Emotinet used the Finite State Automata concept to define the actions triggered emotional responses. One of the participant of SemEval 2007 Task Number 14 [8] used coarse grained and fine grained approaches to identify sentiments in news headlines. They conducted binary classification of emotions in a coarse grained approach and categorized emotions into different levels in fine grained approach. Because of the necessity of a large lexical database, knowledge base approach is found to be difficult. Since social network generates huge amount of data every second, sometimes larger than the size of available lexical database, sentiment analysis became tedious and erroneous.

### 1.3 MACHINE LEARNING TECHNIQUES

[14] Machine learning methods use a training set and a classification evaluation set. Training set contains input feature vectors and their corresponding class labels.Using this training set a classification model is developed that

attempts to classify the vectors of the input function into the corresponding class labels. Then a test set is used to validate the model by predicting the class labels of unseen feature vectors. A number of machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify reviews [9]. Some of the features that can be used for sentiment classification are Term Presence, Term Frequency, negation, n-grams and Part-of-Speech [1]. Such features can be used to determine the semantic orientation of verbs, phrases, sentences and documents. Semantic orientation is the polarity which can either be negative or positive. Domingos et al. [10] found that Naive Bayes works well for certain problems with highly dependent features. This is surprising as the basic assumption of Naive Bayes is that the features are independent. Zhen Niu et al. [11] introduced a new model in which efficient approaches are used for feature selection, weight computation and classification. The new model is based on Bayesian algorithm. Here classifier weights are modified using representative function and special feature. 'Representative function' is the information that reflects a class and the information that helps to differentiate classes is 'Special feature.' We measured the likelihood of each classification using certain weights and thus strengthened the Bayesian algorithm. Barbosa et al. [12] designed a 2-step automatic sentiment analysis method for classifying tweets. When designing classifiers they used a noisy training set to reduce the labeling effort. First they categorized tweets into tweets that were subjective and objective. Subjective tweets are then marked as positive tweets and negative ones. Celikyilmaz et al. [13] developed a pronunciation based word clustering method for normalizing noisy tweets. Words with identical pronunciations are clustered in pronunciation-based word clustering and allocated specific tokens. It have used text processing methods such as assigning numbers to specific tokens, HTML connections, identities of users, and names of target organizations for normaliz

[15] ation. After doing normalization, they used probabilistic models to identify polarity lexicons. With these polarity lexicons as features they conducted classification using the BoosTexter classifier and obtained a reduced error rate. Wu et al.proposed driving probability model for study of twitter sentiments. If @username is contained in a tweet 's body, it affects behavior, and adds to the likelihood impact. Any tweet that begins with @username is a retweet that represents an influenced action and it contributes to influenced probability. They observed that there is a strong correlation between these probabilities. Pak et al. [15] created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they developed a sentiment classifier based on the Naive Bayes multinomial classifier, which uses N-gram and POS tags as features. There is a chance of error in that process because emotions of tweets are classified solely based on the polarity of emoticons in the training collection. The training set is also less effective since there are only tweets with emoticons in it. Xia et al. [16] used an ensemble framework for sentiment classification. Ensemble architecture is obtained by the combination of various feature sets and classification techniques. They used two types of feature sets and three base classifications to shape the ensemble structure in that work. Two types of feature sets are generated using Word-relations and Part-of-Speech knowledge. Naive Bayes, Maximum Entropy, and Vector Support Machines are chosen as the base classifiers. They applied various ensemble methods such as Fixed Combination, Weighted Combination, and Meta-Classifier Combination for Classification of Sentiments and obtained better accuracy. Certain attempts are made by some researches to identify the public opinion about movies, news etc from the twitter posts. V.M. Kiran et al. [17] used information from other publicly available databases such as IMDB and Blippr following appropriate modifications to assist in the analysis of twitter feelings in the film domain.

## 2. PROPOSED SYSTEM

A dataset is created using twitter posts of electronic products. Tweets are short messages with full of slang words and misspellings. So we perform a sentence level sentiment analysis. This is done in three phases. In first phase preprocessing is done. A function vector will then be generated using relevant features. Finally, tweets are categorized into positive and negative classes using various classificators. Based on the number of tweets in each class, the final sentiment is derived.

### 2.1 CREATION OF DATASET

| Dataset | Positive | Negative | Total |
|---|---|---|---|
| Training | 500 | 500 | 1000 |
| Test | 100 | 100 | 200 |

Since standard twitter dataset is not available for electronic products domain, we created a new dataset by collecting tweets over a period of time ranging from April 2013 to May 2013. Tweets are collected automatically using Twitter API and they are manually annotated as positive or negative. Taking 600 positive tweets and 600 negative tweets produces a dataset. Table 1 shows how dataset is split into training set and test set.

## 2.2 PREPROCESSING OF TWEETS

Extraction of keywords in twitter is difficult because of misspellings and slang terms. So to avoid this, a preprocessing step is performed before feature extraction. Preprocessing steps include removing url, avoiding misspellings and slang words. Misspellings are avoided by replacing repeated characters with 2 occurrences. Slang words contribute much to the emotion of a tweet. So they can't be simply removed. Therefore a slang word dictionary is established to replace the slang terms that occur with their corresponding meanings in tweets. Domain knowledge contributes significantly to the development of dictionary slang term.

## 2.3 CREATION OF FEATURE

Vector Feature extraction is done in two steps.Twitter specific features are extracted in the first step. The related basic features of twitter are hashtags and emoticons. The emoticons can be either positive or negative. So they are given different weights. Positive emoticons are given a weight of '1' and negative emoticons are given a weight of '-1'. There may be positive and negative hashtags. Hence the count of positive hashtags and negative hashtags in the function vector was introduced as two separate functions. Not all tweets may have unique features on Twitter. And to get other functionality, a further extraction of the feature must be performed. They are deleted from the tweets after removing twitter unique features. Tweets can be then considered as simple text. Then using unigram approach, tweets are represented as a collection of words. In unigrams, the keywords represent a tweet. We maintain a negative list of keywords, a positive list of keywords and a list of different words which reflect negation. The role vector uses counts of positive and negative keywords in tweets as two separate features. Presence of negation contribute much to the sentiment. So they also add their presence as a related feature. In the presence of several positive and negative keywords all keywords can not be handled equally. Hence a special keyword from all the tweets is chosen. In the case of tweets that either have positive keywords or negative keywords, a search is conducted to find a keyword that has a significant part of the expression. A relevant part of speech is adjective, adverb or verb. Such a relevant part of speech is defined based on their relevance in determining sentiment. Keywords that are adjective, adverb or verb shows more emotion than others. If a relevant part of speech can be determined for a keyword, then that is taken as special keyword. Otherwise a keyword is selected randomly from the available keywords as special keyword. When a tweet includes both positive and negative keywords, we pick every keyword that has a significant part of the expression. When there is a appropriate part of speech for both positive and negative keywords, none of these are selected. If it is positive and '-1' if it is negative and '0' in its absence, special keyword function is assigned a weight of '1'. Part of the speech function is given a '1' value if appropriate, and '0' if not. Thus function vector consists of eight related apps. The 8 features used are a part of speech (pos) tag, special keyword, negation appearance, emoticon, number of positive keywords, number of negative keywords, number of positive hash tags and number of negative hash tags.

## 2.4 SENTIMENT CLASSIFICATION

After creating a feature vector, classification is done using Naive Bayes, Support Vector Machine, Maximum Entropy and Ensemble classifiers and their performances are compared.

## 3. CLASSIFICATION TECHNIQUE

There are different types of classifiers that are generally used for text classification which can be also used for twitter sentiment classification.

## 3.1 NAIVE BAYES CLASSIFIER

Naive Bayes Classifier makes use of all the features in the feature vector and analyzes them individually as they are equally independent of each other. The conditional probability for Naive Bayes can be defined as

$$P(X|yj ) = \Pi m\ i=1P(xi\ |yj )$$

'X' is the feature vector defined as X={x1,x2,....xm} and yj is the class label. Here, there are numerous independent features in our work such as emoticons, emotional keywords, count of positive and negative keywords and count of positive and negative hashtags that are effectively used for classification by Naive Bayes classifier. Naive Bayes does not find the interpersonal relationships. So it can't use the relationships between speech tag part, emotional keyword and negation part.

## 3.2 SVM CLASSIFIER

SVM Classifier uses large margin for classification. It separates the tweets using a hyper plane. SVM uses the a discriminative function defined as

$$g(X) = w\ T\ \varphi(X) + b$$

'X' is the vector of the function, 'w' the vector of weights and 'b' the vector of bias. φ() is the non linear mapping from input space to high dimensional feature space. 'w' and 'b' are learned automatically on the training set. Here we used a linear kernel for classification. It maintains a wide gap between two classes.

## 3.3 MAXIMUM ENTROPY CLASSIFIER

No conclusions about the relationship between features are made in Maximum Entropy Classifier. This classifier also tries to optimize machine entropy by estimating class mark conditional distribution.The conditional distribution is defined as

$$P\lambda(y|X) = 1/Z(X)\exp\ (X\ i\ \lambda i fi(X, y)\ )(3)$$

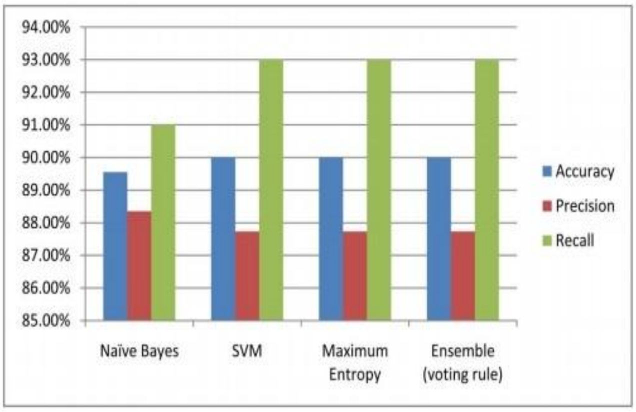
The function vector is 'X' and the class identifier is 'y.' Z(X) is the normalization factor and λi is the weight coefficient. fi(X, y) is a function that is set to

$$fi(X, y) = \{1,\ X=xi \text{ and } y = yi\ (4)$$

0, otherwise

The relationships between part of speech tag, emotional keyword, and negation are effectively used for classification in our function vector.

## 4. ENSEMBLE CLASSIFIER




Ensemble classifiers can be of different types. They try to make use of the features of all the base classifiers to do the best classification. The base classifiers used here are Naive Bayes, Maximum entropy and SVM. Here an ensemble classifier is generated by voting rule. The classifier will be graded based on the most classifier performance.

Using Twitter API, tweets related to products are collected. A dataset is created using 1200 twitter posts of electronic products. Dataset is split in to a training set of 1000 tweets and a test set of 200 tweets. We used Stanford postagger1 for extracting part of speech tag from tweets. Since we have selected product domain, there is no need of analyzing subjective and objective tweets separately. To identify the quality of product, both of these qualities contribute similarly. This shows how context or domain information affects sentiment analysis. These classifiers are tested using Matlab simulator. We used three types of basic classifiers (SVM, Nave Bayes, Maximum Entropy) and ensemble classifier for sentiment classification. SVM and Naive Bayes classifiers are implemented using Matlab built in functions. Maximum Entropy classifier is implemented using MaxEnt software2 . Performance of these classifiers is shown in Fig. 1. All these classifiers have almost similar performance. Naive Bayes has better precision compared to the other three classifiers, but slightly lower accuracy and recall. SVM, Maximum Entropy Classifier and Ensemble classifiers have similar accuracy, precision and recall. They obtained an accuracy of 90% whereas NaiveBayes has 89.5%. This shows the quality of the feature vector selected for the product domain. This function vector helps to evaluate better feelings given the selected classifier.

## 5. CONCLUSIONS

There are different techniques of Symbolic and Machine Learning to classify text sentiments. Techniques of machine learning are simpler and more effective than the Symbolic techniques. These methods can be used for examination of feelings on twitter. There are other challenges when it comes to recognizing emotional keywords from tweets that have several keywords. It is also difficult to handle misspellings and slang words. To deal with these issues, an efficient feature vector is created by doing feature extraction in two steps after proper preprocessing. Twitter specific features are extracted and added to the function vector in the first step. After that, these features are removed from tweets and again feature extraction is done as if it is done on normal text. These features are also added to the feature vector. Classification accuracy of the feature vector is tested using different classifiers like Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers. All these classifiers have almost similar accuracy for the new feature vector. For the domain of electronic goods this function vector performs well.

## REFERENCES

1. Y. Mejova, “Sentiment analysis: An overview,” Comprehensive exam paper, available on http://www. cs.

2. E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, “Automatic sentiment analysis in on-line text,” in Proceedings of the 11th International Conference on Electronic Publishing, pp. 349–360, 2007.

3. P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

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

5. C. Fellbaum, “Wordnet: An electronic lexical database (language, speech, and communication),” 1998.

6. D. Pucci, M. Baroni, F. Cutugno, and A. Lenci, "Unsupervised lexical substitution with a word space model," in Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence, Citeseer, 2009.

7. A. Balahur, J. M. Hermida, and A. Montoyo, "Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model," Affective Computing, IEEE Transactions on, vol. 3, no. 1, pp. 88–101, 2012.

8. C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in Proceedings of the 4th International Workshop on Semantic Evaluations, pp. 70–74, Association for Computational Linguistics, 2007.

9. G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," International Journal, vol. 2, no. 6, 2012.

10. P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1997.

11. Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for microblog by machine learning," in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, 2012.

12. L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36–44, Association for Computational Linguistics, 2010.

13. A. Celikyilmaz, D. Hakkani-Tur, and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 79–84, IEEE, 2010.

14. Y. Wu and F. Ren, "Learning sentimental influence in twitter," in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119–122, IEEE, 2011.

15. A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010, 2010.

16. R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

17. V. M. K. Peddinti and P. Chintalapoodi, "Domain adaptation in sentiment analysis of twitter," in Analyzing Microtext Workshop, AAAI, 2011.