Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 6– No.9, April 2017 – www.caeaccess.org

Sentiment Analysis of Twitter Feeds using Machine Learning, Effect of Feature Hash Bit Size

Silas Kwabla Gah
University of Ghana
Department of Computer Science
Legon. Ghana

Nana Kwame Gyamfi
Kumasi Technical University, Kumasi,Ghana
Department of Computer Science

Ferdinard Katsriku, PhD
University of Ghana
Department of Computer Science
Legon, Ghana

ABSTRACT
Sentiment Analysis is a way of considering and grouping of opinions or views expressed in a text. In this age when social media technologies are generating vast amounts of data in the form of tweets, Facebook comments, blog posts, and Instagram comments, sentiment analysis of these user-generated data provides very useful feedback. Since it is undisputable facts that twitter sentiment analysis has become an effective way in determining public sentiment about a certain topic product or issue. Thus, a lot of research have been ongoing in recent years to build efficient models for sentiment classification accuracy and precision. In this work, we analyse twitter data using support vector machine algorithm to classify tweets into positive, negative and neutral sentiments. This research try to find the relationship between feature hash bit size and the accuracy and precision of the model that is generated. We measure the effect of varying the feature has bit size on the accuracy and precision of the model. The research showed that as the feature hash bit size increases at a certain point the accuracy and precision value started decreasing with increase in the feature hash bit size.

General Terms
Hadoop, Data Processing, Machine learning

Keywords
Sentiment Analysis; Machine Learning; Support Vector Machine; Feature Hashing

1. INTRODUCTION
Machine learning involves the study of algorithms that will automatically improve its performance as they learn from the training dataset. This way models are designed to learn from data and been able to predict future results based on the past data supplied to the model. Machine learning is a discipline of computer science or type of artificial intelligence which provides computers with the ability to learn without being explicitly programmed. It focuses on the development of computer programs that can teach themselves to grow and change when exposed to new dataset.

The Internet has brought about a vast change in the way people express their views and opinion. It is mostly now done through tweets, Facebook comments, blog posts, and Instagram comments etc. Organizations and institutions greatly depend on these users generated content to make informed decisions. This has attracted a lot of researchers to work in this area of text classification. This area of machine learning is very important and a well-studied area with a wide variety of application in this day. For accurate performance of a classifier as good feature selection is of paramount importance. Features or attributes are the individual characteristics that combined together to make a training dataset. The ability to choose the right and relevant features out of the other features in a large set of data is particularly paramount for accurate text classification. This is the motivation of this research work to find out the feature size that helps in getting accurate and precise result for the classifier.

There is a growing interest in the area of machine learning due its widespread application. Much of this research effort has been aimed at finding out the best and the most reliable way to carry out sentiment analysis on text data from large sources such as twitter or Facebook which is used by a lot of organization to judge how their product or service offered is doing. A lot of research work has gone into sentiment analysis of twitter data but the measure of the accuracy of the model against the feature hash size used has not been investigated much. This is what this research works seeks to investigate it.

Many applications of the twitter sentiment analysis have been carried out, researchers like [1] have used twitter sentiment analysis to find out if one could answer questions such as "Is it possible to predict stock prices of a company from public sentiment? Is it possible that it would be easier to predict stock prices of certain companies when compared to other companies?". This prediction will be extremely impossible considering the number of tweets and the fast nature these tweets will be coming in. The author [2] have come out with a mining technique to predict the stock prices of 30 NASDAQ listed companies by analyzing 15 million tweets.

In this work, the researcher gathered some works already carried out and critically examines the different technologies, algorithms, and tools that facilitate modeling such systems.

2. RELATED WORK
In recent years of lot of work has been done in the field of sentiment analysis by researchers in the field of computer

science. Much of this research has attempted to find out the best and the most reliable way to carry out sentiment analysis on text data from source such as twitter or Facebook which is used by a lot organization to get the views of their product or services offered to customers.

In order to determine the sentiment analysis of a source text data, there are two main methodologies in achieving this aim. These are the symbolic technique and the machine learning approach [3]. The symbolic technique involves using a word dictionary to compare with tweets. As the word dictionary

have each word annotated with polarity, comparing with tweets messages one can find the polarity of each word in a tweet. Machine learning employees the services of algorithm to study and make informed decision on the polarity of tweets.

2.1 Symbolic Technique

Most of the research in unsupervised sentiment classification using symbolic techniques which makes use of available lexical resources. The lexical resource is a database consisting of one or several dictionaries, depending on the number of languages that is being addressed. It could be bilingual or multilingual. In [4] used a bag of words approach for sentiment analysis. In this work, the authors did not consider the relationship between the words rather a document is represented as a mere collection of words. In arriving at the overall sentiment, individual words sentiment are aggregated to obtain a sentiment value. The researcher found the polarity of a review based on the average sentiment position of tuples extracted from the review where tuples are phrases having adjective or adverbs.

In a different study [5] the author employed the services of the lexical database WordNet [6] is used to determine the emotional content of a word along different dimensions. WordNet consists of words connected by synonym relations.

The authors of [7] used a lexicon-based approach to analyze the data which was stored on the Hadoop platform. The research was done on a single node but the approach showed that the response from the sentimental model was fast and some how efficient, but it could be more efficient if multiple nodes are applied on the Hadoop.

Knowledge base approach is found to be difficult due to the requirement of a huge lexical database which must be kept and referenced every now and then during the experiement. Due to the fact that all social media networks can generate huge amount of data every second [8], the lexical database can not handle such requirments and becomes erroneous and tedious work to be carried out.

2.2 Machine Learning Techniques

The field of machine learning seeks to provide computers with the ability to learn without being explicitly programmed. It focuses on the development of computer programs that can trach themselves to grow and take decisions when exposed to new set of data. Sentiment analysis is a special case of text mining that is increasingly important in business intelligence and social media analysis.

Machine learning technique approached employed the services of a training dataset and a test dataset for the classification purposes. This way the training dataset contains a number of feature vectors and an equivalent class labels. Employing the services of the training set, a classification model is developed which is used to classify the input feature vector into its various class labels. After this proceeding the test set is used to validate the model by predicting the class labels of inkown feature vector. Support Vector Machine is a class of supervised machine learning technique, used to train a sentiment classifier by taking the frequency of occurrence of various text contained in a text source or tweets [9].

There are quite a number of machine learning techniques out there such as Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM) which are used to perform this classification of text and reviews by these authors [10]. Even though it was found out that all these technique could handle large text data source,it was extablished that the SVM model yeilds better and much accurate result compared to the other models.

In [11] the researchers looked at how varying the training set size on the classification impacts on accuracy and F-score of SVM and Naïve Bayes classifiers. In the research it was concluded that the SVM accuracy largely surpassed that of the Naïve Bayes classifier hence the need to opt for the SVM model for the training the model.

Domingos, [12] found that Naïve Bayes works very well in certain situations with high dependent features. The finding by the author is surprising as the basic assumption of Naïve Bayes is that the features are independent.The features dependent is the research by Domingos clearly established the fact that the accuracy and precision of a model highly depends on the features. Zhen Niu [13] came out with a new model in which efficient approaches are used for the feature selections, weight compuation and classification, this new model was used in the Bayesian algorithm. Here weights of the model are adjusted by making use of representative feature and unique feature. Representative feature is the information that represents a class and unique feature is the information that helps in distinguishing classes. Using those weights, calculated the probability of each classification and hence improved the Bayesian algorithm.

Barbosa, [14] developed a 2-step automatic sentiment classifier for classification of tweets. They employed the services of noisy training set to reduce the labelling effort in developing classifiers. First step was to calssify the tweets into subjecive and objective tweets. The second step was subjective tweets are classified as positive and negative tweets.

Xia [15] used an ensemble framework for sentiment classification. In this Ensemble framework obtained by combining various feature sets and classification technique. This work used two types of feature sets and three base classifiers form the ensemble framework. Two types of features sets were created using Part-of-speech information and Word-relations. Naïve Bayes, Maximum Entropy and Support Vecor Machines were selected as base classifers. They then applied different ensemble methods such as the Fixed combination, Weighted combination and Meta-classifer combination for the sentiment classification and obtained a better accuracy.

Pak [16] created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating using emoticons. By the use of this corpus, the researchers built a sentiment classifer based on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features. This introduced a margine of error since emoticons of tweets in training set are labeled soley based on the polarity of emoticons. The training set is also less efficieint since it contains only tweets having emoticons.

Twitter [17] alluded to the difficulty in dealing with sentiment analysis of twitter data due to slang words, misspelling and the different domain contexts that language meaning may have. The researcher compared different machine learning techniques such as Support Vector Machine

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 6– No.9, April 2017 – www.caeaccess.org

(SVM), Naïve Bayes Classifier and they came to the conclusion that all the techniques were simpler and more efficient than symbolic techniques.

In this research [18], it was found that the Hadoop cluster of servers which divides data among the various nodes, hence making the processing of data faster was used in collecting and classifying of tweets. This research [19] looked into the effective sentiment analysis on twitter data by employing the Apache Flume and Hive. Hive are technologies built on top pf the Hadoop which makes querying and structuring of one data set easy, faster and more efficient. The reseachers employed a real-time stream of data using flume, which enables the real-time inflow of tweets.

3. MATERIAL AND METHOD

In the current work approach of trying to analyze twitter dataset, we used the machine learning technique with support vector machine model. This is a form of supervised machine learning where the model is trained before test data is applied to it. In this approach of trying to analyze twitter dataset, we used the feature hashing extraction technique. The researcher focused on a framework where the pre-processor stage is applied to the raw sentences which makes it more appropriate to understand the data better. Further down the stage the support vector machine technique is employed to train the dataset with feature vectors. Identifying the common characteristics of a set of objects that are representative of their class is of great interest in classifying the object. The main objective of text classification is to assign some piece of text to one or more predefined classes or categories. This text could be anything such as documents, news articles, or a tweet.

The researcher focused on a framework where the pre-processor stage is applied to the raw sentences which make it more appropriate to understand the data better. Further down the stage, the support vector machine technique is employed to train the dataset with feature vectors. The feature hash bit size is varied incrementally and the accuracy and precision of the model is measured and examined. The complete description of the approach is given and explained in the sub sections and the block diagram as shown in in Figure. 1

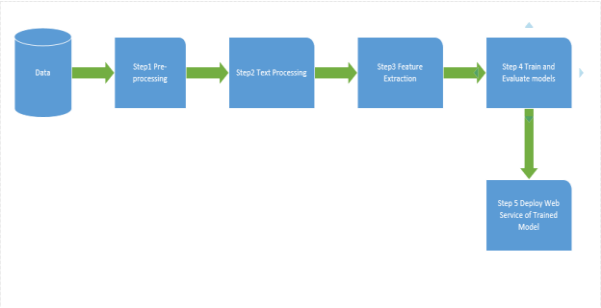

Fig 1: Workflow for experiment

It is important that the various steps in the workflow follow accordingly, one after the other, this way it is assured that one output serves as an input to the next step. The first four steps in the Figure 1, represent the phase where the text classification is done. During this phase, the text instances are loaded into the machine learning experiment stage and the text is cleaned and filtered. Different types of numerical features are extracted from the cleaned text and the models are trained on different feature types. The last stage is where the performance of the trained model is evaluated on unknown or unseen text instances and the best models determined based on a number of evaluation criteria. The various steps in the workflow is explained in the following section.

3.1 Data Processing

During this phase of the experiment, the data is loaded from the twitter source, this is recollected tweets. The reader allows data to be added from various sources into the experiment stage. The Project column module is used to select the needed column which has the features the researcher is interested in, and only these columns are selected during execution. The metadata editor module is used to rename the various columns as label column and text column respectively. The next module at this phase is the clean missing values module. This module will clean and replace all missing values with empty strings. The project column is again called to select the text column fields. The experiment is run at this stage to ensure that all is correct up to this stage.

In this experiment two fields are used for training the model; these are the tweet text field and sentiment label assigned to the tweet. This is shown in Figure 2.

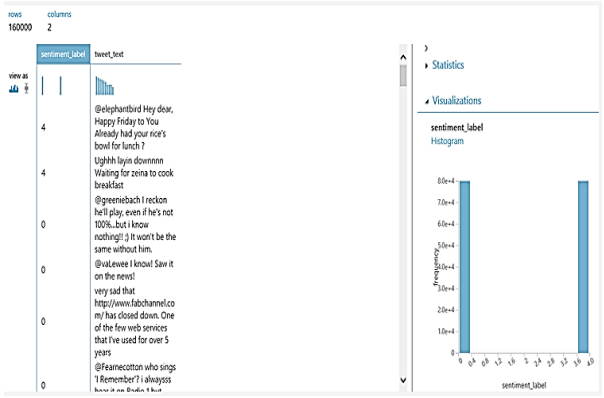

Fig 2: Fields for training model

3.2 Text Preprocessing

Due to the largely unstructured nature of the tweets, it will be good to preprocess the data before it can be analyzed. This is where the power of the R language is employed to help streamline and remove unnecessary data from the tweets. Characters like punctuation marks, special characters and digits are taken out from each tweet. Also, case normalization is done on the tweets to bring all to one case tense, that is upper cases to lower case. One important task carried out at this stage is the replacing of special characters with spaces. Also, the removal of duplicate characters is done at this phase, as well as the conversion of all uppercase letters ones. Stem words are converted, stemming is the process of reducing inflected words to their word stem, base or root form.

3.3 Feature Hashing

The main goal of feature hashing is to convert variable-length unstructured text data into equal-length numeric feature vectors. The added benefit of the hashing feature is that it reduces the dimensionality of the data and makes a lookup of feature faster by replacing string compression with hash value comparison. For this experiment, the hashing bit is set to 17 and the number of N-grams to 2. This sets the hash table to 217 or 131,072 entries in which each hashing feature will represent one or more unigram or bigram features. The

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 6– No.9, April 2017 – www.caeaccess.org

classification time and complexity of a trained model depends on the number of features, that is the dimensionality of the input space. The bit of the hashing was increased to 20 to see the significant of this in the model accuracy. For text classification tasks, the number of features resulting from feature extraction is high because each n-gram is mapped to a feature. The filter-based feature selection module is used to select a more compact feature subset from the exhaustive list of extracted hashing features. The aim is to avoid the effects of dimensionality and to reduce the computational complexity without harming the classification accuracy. The split module in the experiment is used to split the data into train and test sets where the split is stratified. The stratification will maintain the class ratios in the two output groups. The first 80% of the Sentiment140 sample tweets, a publicly available dataset created by three graduate students at Stanford University: Alec Go, Richa Bhayani, and Lei Huang is used for training and the remaining 20% for testing the performance of the trained model.

3.4 Train Prediction Model

This phase uses the Support Vector Machine algorithm to train the model. There is a number of learning algorithms that can be used to train the module, but the SVM is used in this experiment. Support vector machines (SVM) are supervised learning models that analyze data and recognize patterns. They can be used for classification and regression tasks. The classifier that is created from this module is useful for predicting between two possible outcomes that depend on continuous variables. The random seed is set to 123 and random iteration of the algorithm is set to 10 times.

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

$$g(X) = w^T \emptyset(X) + b$$

'X' is the feature vector, 'w' is the weights vector and 'b' is the bias vector. () is the nonlinear mapping from input space to high dimensional feature space. 'w' and 'b' are learned automatically on the training set.

3.5 Model Evaluation

In order to evaluate the generalization ability of the trained Support Vector Machine model on unseen data, the output model and the test data set are connected to the scoring model in order to score the tweets of the test set. This is then connected to evaluate module to get a number of performance metrics. The table below shows the confusion matrix which is used to assist in the calculation of the evaluation of the model.

Table 1. Confusion matrix

	Predicted Positives	**Predicted Negatives**
Actual Positives	Number of true positives (TP)	Number of false negatives (FN)
Actual Negatives	Number of false positives (FP)	Number of true negatives (TN)

The formula that is used as a reference point in assessing the various model is shown below:

Accuracy: This is the value of the entire true predicted value against all the predicated value. This formula can be calculated as show below

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

Precision: This is the value of true positive prediction against all positive prediction. This is calculated as shown below

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

Recall: This is the value of true positive prediction against all actual positive. The formula can be calculated as shown below:

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

4. RESULT

The experiment was carried out with a dataset of 1,600,000 automatically annotated tweets. Out of this quantity, 1,280,000 tweets were used in training the model and the remaining 320,000 tweets used to test the model.

The dataset visualization is shown in the figure below. Figure 3 shows the tweets used and the sentiment labels for each of them.

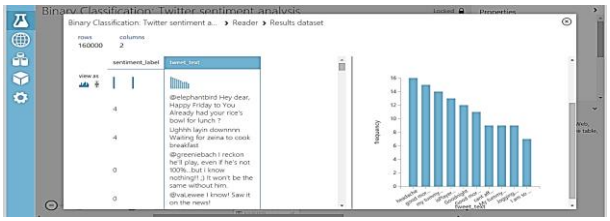

Fig 3: Data visualization

The model has two sets of data as mentioned earlier. The trained dataset yielded the graph in figure 4. The precision of this trained model was 0.855. The closer the line is to the 1 vertical axis, the better the model. The results also show true positive and true negative values.

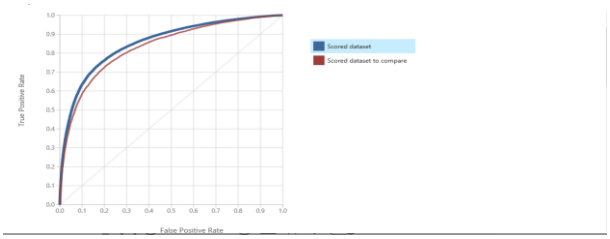

Fig 4: ROC/AUC for Feature Hashing bit 17

Table 2. Values of Feature Hash bit at 17

Precision	**TP**	**FN**	**FP**	**TP**	**Accuracy**
0.763	**51339**	**12661**	**15980**	**48020**	**0.776**

The most interested part of the result is the AUC, which records the proposition of the correctly classified value of the classifier. In addition to these metrics are the True positive values (TP). True positive values consist of the true positive (TP) and the true negative (TN). Similarly, the incorrectly

classified instances are called false positives (FP) and false negatives (FN).

In addition to these metrics are questions such as how many were classified correctly, which is called precision of the model.

TP / (TP + FP)

There is also the question of how many the classifier classified correctly (TP). This is the recall of the classifier

TP / (TP + FN)

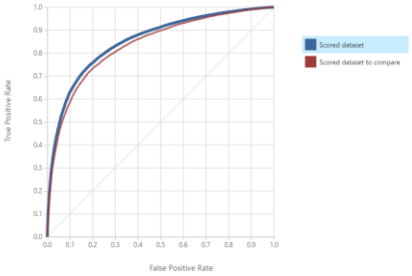

Fig 4: ROC/AUC for Feature Hashing bit 20

The result for the feature hashing when the bit size was set to 20 is shown in figure 5. The precision and true positive and true negative values of the test data is shown below.

Table 2. Values of Feature Hash bit at 15

Precision	TP	FN	FP	TP	Accuracy
0.765	51337	12663	15811	48189	0.778

When the featuring hashing bit size was increased to 20 the following resulted in the precision value for the trained data was obtained.

The precision of the feature hashing when the bit size was increased to 20 bits was lower than when the feature hashing bit was 17.

From table 3 the precision value for the feature hashing bit 20 is lower than that of the feature hashing bit 17. However, the accuracy of the feature hashing 20 is higher than that of feature hashing bit 17.

Table 3. Values of Feature Hash bit at 16

Precision	TP	FN	FP	TN	Accuracy
0.766	51417	12583	15685	48315	0.779

Table 4. Values of Feature Hash bit at 18

Precision	TP	FN	FP	TP	Accuracy
0.761	51279	12721	16083	47917	0.775

Table 5. Values of Feature Hash bit at 19

Precision	TP	FN	FP	TP	Accuracy
0.761	51261	12739	16098	47902	0.775

Table 6. Values of Feature Hash bit at 20

Precision	TP	FN	FP	TP	Accuracy
0.760	51221	12779	16200	47800	0.774

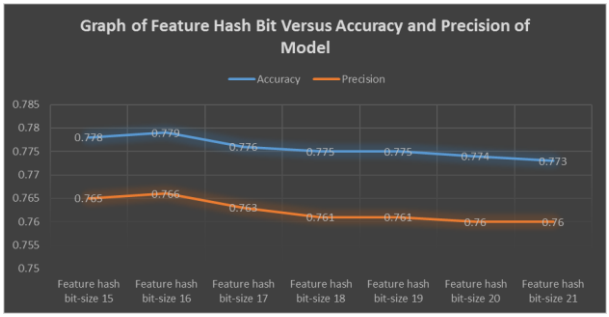

Fig 4: Graph Feature hash bit values versus Accuracy and Precision.

From the graph, it is clear that the accuracy and precision of the model decreases as the feature hash bit size is increased. Even though theoretically it would be more sensible for the model to have higher accuracy if the feature hash bit size is higher since there would be more values for classification. But it is clear that is not the case.

5. CONCLUSION

There are a number of different symbolic and machine learning techniques to identify sentiments from Twitter data. Machine learning techniques are simpler and efficient when it comes to large datasets. These techniques can be applied to Twitter sentiment analysis.

Moreover, the results show that the increase or decrease in the feature hashing did not have much significant difference in the accuracy and precision of the model that was created.

It is clear that when the feature hash bit size was increased the accuracy and precision of the model decreased and not greater increase in these values. It can be concluded that not necessarily higher values of feature hash bit size will yield higher value of accuracy and precision.

It was also observed that when the feature hashing was increased to a higher level more memory was required for the

experiment to be carried out. This made the processing time of the model higher than lower values of the feature hash bit size.

When the values were higher processing time was high.

There are however some issues when it comes to dealing with such large datasets. It was observed that the trained data set gives a higher accuracy than the test data. This was due to the large data set that the machine used and learned from.

From the research, it was observed that the result was much more accurate when the text was composed of more words that single and short words. The one critical factors that affect the accuracy of the model is the feature hashing technique that is used. The more space that is generated the more accurate the model since more storage and text to compare with. The side effect the time is taken to do the search on these feature vectors. Even though the greater the hashing bit is, one needs to be mindful of the time it takes to search through all entries to enable fast retrieval of record. It was observed that when the feature hash bit was increased the precision of the model rather decreased.

REFERENCES

[1] K. C. C. C. a. C. O. Li Bing, "Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movement," in *14 Proceedings of the 2014 IEEE 11th International Conference on e-Business Enginerring*, 2014.

[2] K. C. C. C. a. C. O. Li Bing, "Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movement," in *14 Proceedings of the 2014 IEEE 11th International Conference on e-Business Enginerring*, 2014.

[3] B. P. K. M.-F. E., "Automatic Sentiment Analysis in On-line text," *ELPUB*, pp. 349-360, 2007.

[4] P. D. Turney, "Thumbs up or thumbs down?:semantic orientation applied to unsupervised classification of reviews," *40th annual meeting on association for computationa; linguistics*, p. 417–424, 2002.

[5] K. M. M. R. J. M. a. M. D. R. J., Using wordnet to measure semantic orientations of adjectives, 2004.

[6] C. Fellbaum, "Wordnet: An electronic lexical database," 1998.

[7] C. Kaushik and A. Mishra, "A scalable, Lexicon based technique for sentiment analysis," *Journal of International Journal in Foundation of Computer Science & Technology (IJFCST)*, 2014.

[8] T. WebSite, 1 March 2015. [Online]. Available: https://about.twitter.com/company.

[9] B. M. D. S. T. Croft, Search Engines: Information Retrieval in Practice, Addison Wesly Publishing Company, 2009.

[10] G. V. R. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," *International Journal 2*, vol. 2, p. 6, 2012.

[11] O. e. a. Abdelwahab, "Effect of training set size on SVM and Naive Bayes for Twitter sentiment analysis," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2015.

[12] P. D. a. M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," in *Machine Learning*, 1997, pp. 103-130.

[13] Z. Y. a. X. K. Z. Niu, "Sentiment classification for microblog by machine learning in Computational and Information Sciences," *IEEE*, pp. 286-289, 2012.

[14] L. B. a. J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *23rd International Conference on Computational Linguistics*, 2010.

[15] C. Z. a. S. L. R. Xia, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences: an International Journal*, vol. 181, no. 6, pp. 113-1152, 2011.

[16] A. P. a. P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of LREC*, 2010.

[17] N. M. R. R. S, "Sentiment Analysis in Twitter using Machine Learning Techniques," *IEEE*, 2013.

[18] S. B. Y. Mane and S. S. V. Kazi, "Real Time Sentiment Analysis of Twitter Data Using Hadoop," *International Journal of Computer Science and Information Technology*, 2014.

[19] G. M. S. B. C. Penchalaiah, "Effective Sentiment Analysis on Twitter Data using Apache Flume and Hive.," *International Journal of Innovative Science*, 2014.

[20] O. e. a. Abdelwahab, "Effect of training set size on SVM and Naive Bayes for Twitter sentiment analysis," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2015.