Social Media Sentiment Analysis using Machine Learning and Optimization Techniques

E. M. Badr
Scientific Computing Department
Faculty of Computers and Informatics
Benha University, Egypt

Mustafa Abdul Salam
Scientific Computing Department
Faculty of Computers and Informatics
Benha University, Egypt

Mahmoud Ali
Scientific Computing Department
Faculty of Computers and Informatics
Benha University, Egypt

Hagar Ahmed
Scientific Computing Department
Faculty of Computers and Informatics
Benha University, Egypt

ABSTRACT
Recently, there are emergence and advent of data Inter-personal interaction web sites, micro blogs, wikis, in addition to Web applications and data, e.g. tweets and web-postings express views and opinions on different topics, issues and events in many applications, in addition to, different domains that includes business, economy, politics, sociology, and etc., which are resulted from offering immense opportunities for studying and analyzing human views and sentiment. The objective of sentiment analysis is to classify a speaker's or a writer's attitude towards various events or topics and arranging data into positive, negative or neutral categories. Sentiment analysis means determining the views of a user from the textual content regarding that topic i.e. how one feels about it. It might be used to classify the text content. Various researchers have used a widespread sort of methods to teach the classifiers for the Twitter dataset with various results. The research uses a hybrid method of using Swarm Intelligence optimization algorithms with classifiers. For each tweet, pre-processing will be done by performing various processes i.e. Tokenization; removal of stop-words and emoticons; stemming. Then their feature vectors are being made by the calculation of TF-IDF and optimized with (PSO) and (ACO) before performing the binary text categorization. Naïve Bayes and Support Vector Machine may be those machine learning technicalities used for the binary classification of tweets.

Keywords
SVM, Naïve Bayes, ACO, PSO, Sentiment analysis, Twitter

1. INTRODUCTION
"Sentiment analysis is particular case class for computational strategies which naturally excerpts and epitomize those conclusions from opinions such monstrous quantity of information which the average person reader is not able to process" [10]. In the era of advanced Internet, Technology and Individuals all over the world be part of to each other's through social media web sites such Face book, Google+, Twitter, Instagram, etc. Twitter has nearly 300 million active users with Feature Optimization; it's been performed using millions of tweets for every day; which makes it a main social media internet site worldwide. As Twitter has this huge range of users and enormous data, it has usually been used as "an informative resource by various organizations to research public opinions and gather critical feedback" [11]. In a tweet, users can write their views regarding any topic or general thoughts in a maximum length of **one hundred forty** characters only. Due to this limited tweet length, people write in a very concise manner by using slangs; which makes sentiment analysis a hard task. Sentiment Analysis may be characterized, similarly, as the process about categorizing the opinions expressed through tweet to understand the user views about that topic. "It is beneficial for the marketers to examine and analyze the opinions of the public towards their brand and existing/newly released products; which would help them to evaluate their performance and improve it" [12].

Swarm Intelligence Optimization Techniques are used which are inspired by the behavior technique of the group of insects in our nature. ACO techniques inhibit the natural behavior technique of ants which aims to discover the food through that most brief path to their colony. They keep the track of each path through the deposition of the pheromones; which eventually evaporates; while returning to the colony. A shorter path will have higher pheromone density as it's been followed very frequently. The same behavior has been integrated in the computer artificial ants to search the best solutions for a given problem. PSO inhibits the social behavior of birds as they intend to live in flocks. All the birds try to discover the food in an area. But they realize how a long road the food is. This ends to the tracking of birds that are near the food. This behavior has been integrated to improve the candidate solution iteratively locally and globally which helps to will find out those best optimized solution.

Machine Learning techniques would use to categories the tweets into positive or negative classify. Results can be obtained somewhat better by way of the usage of using the supervised machine learning model. NB and SVM are two learning methods considered for this research. Besides, we will make fruitful suggestions for use of diverse algorithms to different classes of social net data. The model has been carried out in java and then tested against tweets and its performance has been evaluated by considering four parameters: Accuracy, Precision, Recall and F-Score.

2. RELATED WORK
Sentiment Analysis is the thorough research of how opinions and Perspectives can be related to one's emotion and attitude shows in natural language respect to an event. Recent events show that the sentiment analysis has reached up to great achievement which can surpass the positive vs. negative and deal with whole arena of behavior and emotions for different communities and topics. Inside the subject of sentiment analysis using different techniques appropriate quantity of studies has been completed for prediction of social opinions. *Ankita Gupta et al.* [1] In this paper, SVM and KNN based hybrid model is presented to improve the classification accuracy. The proposed method classified the tweets in positive, negative and neutral sentiments whereas much of the literature in this field is associated with 2-way classification.

The work of proposed model has gone through preprocessing stage, features generation stage and classifiers learning stage. The analytical evaluation of proposed model is done in terms of accuracy and f-measure. The comparative observations are taken against the SVM and KNN methods. The comparative results show that the proposed model has improved the accuracy and f-measure of tweet class prediction. *Bharat Naiknaware* et al. [2] if the MAE is smaller than accuracy. The results show that the performance of the classifiers is same. There is marginal difference in the MAE. The performance of the classifiers was made for seven datasets (Budget2017, Demonetization, GST2017, Digital India, Kashmir, Make in India, Startup India). In the Budget2017 dataset Naïve Bayes performs best, In Demonetization dataset Naïve Bayes performs best. In the GST2017 SVM is showing best performance, whereas in the Digital India, Kashmir, Make in India and Startup shows Max Entropy performs best. Here, we also find that the Mean Error for predicting the Mean Absolute Error easily. *Christianini and Taylor*. [3] Published and shared the knowledge about SVM which is machine learning algorithm. The authors manage to give deep understanding about algorithm and how to approach the SVM algorithm in order to implement it to solve the practical problems. The approach will be theoretical as when the book was published, the research was on going on every field. *Malhar and Ram* [4] employed supervised machine learning techniques and artificial neural networks to classify twitter data along with case study of Presidential and Assembly elections which results SVM outperforms all other classifiers. The authors proposed a methodology to predict the outcome of election results by utilizing the user influence factor. To carry out reduction in dimension the authors combined the Principle Component Analysis with SVM. *Martineau and Finin*. [5] Proposed a technique called Delta TFIDF which measure word scores efficiently before classification. Delta TFIDF was easy to understand, implement and compute. For sentiment classification the authors used support vector machines to achieve better accuracy with Delta TFIDF and using data sets of movie reviews. The authors said that Delta TFIDF is better than TDFIF feature and count term raw for all sizes of documents that weights for congressional detecting support for bill, sentiment polarity classification and subjectivity detection. The authors stated Delta TFIDF is first measuring approach to boost and identify the relevance of selective words using the calculated unsupervised distribution of features before classification between the two classes. *Mohammad et al.* [6] developed two SVM classifiers, one is term level task which determines the sentiment of a word in the message and one is message level task which determines the sentiment of messages such as SMS and tweets. The authors took part in a competition where 44 teams came in their submissions stood fist in work on tweets, getting 88.93 F-core in term-level task and 69.02 F-score in message-level task. The authors executed sentiment, semantic and surface-form features. The authors also produced two big term-sentiment associations, first with emoticons from tweets and second with sentiment –term hashtags from tweets. *Neri et al.* [7] performed sentiment analysis on newscast over more than 1000 Facebook posts and then compared the sentiment for dynamic company La7 and Rai – the Italian social broadcasting company which is emerging company. The authors observations were mapped with the study conducted by the Italian research institute highly specialized in study of media at empirical and theoretical level, occupied in the study of communication of politics in the mass media known as Osservatorio di Pavia. The authors experiment done by Knowledge Mining System which is used by security related agencies and institution of government in Italy to control information contained Web Mining and OSINT. *Pablo et al.* [8] provided versions of Naive Bayes classifiers for identifying polarity about English tweets. Two distinct versions of (NB) classifiers had been constructed particularly Baseline (educated to categorise tweets as positive and tweets as negative, neutral), and Binary (uses a polarity lexicon and classifies tweets as positive and tweets as negative. Those Characteristics recognized by means of classifiers were taken from (noun and verb and adjectives Also adverb). Multi phrases from special resources and Valence Shifters. *Po-Wei Liang et.al* [9] proposed a system in U.S. elections 2012 for presidential candidates using real-time evaluation of sentiment on online microblogging site twitter. In order to collect the poll data, the traditional analysis of election takes much time, but with the help of this system it takes data from more people with help of twitter, a microblogging service. It helps the social people like scholars, media and politician to broadcast their future perspective of the public opinion and electoral process. The authors finally concluded that the system and approach are generic, and should be adopted easily and spread across various other domains.

3. METHODOLOGY

So as will perform sentiment analysis, would needed with gather data from the desired source (here Twitter). This data undergoes various steps of pre-processing which makes it more machine sensible than its previous form.

One of the best things that happen on machine learning is that the algorithms can memorize the data and when we need to use it for another data it has a poor performance, this behavior is called over fit. To avoid this problem, I work with test driven methodology. Each dataset is divided in three random parts and each part in three more divisions:

- Train (60%): Used into feed the machine learning in algorithm on the learning process.
- Test (20%): Used to see if our algorithm is over fitting or not.
- Validation (20%): Used to assess the execution of the algorithm**.**

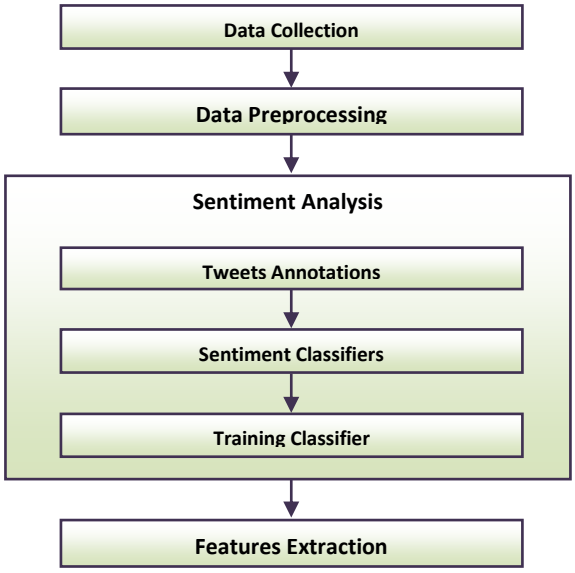

Fig. (1) : General Methodology for sentiment analysis

3.1 Data Collection

Is concerned with the correct acquisition of data; regardless of the various methods depending on the field, the accentuation for guaranteeing accuracy stays those same. The number one aim of any data collection attempt is to obtain quality data that may be easily translated to analyze rich data evaluation which can result in dependable and conclusive answers to questions that need been posed.

So, it is Tweet collection involves gathering relevant tweets about the particular area of interest. The tweets are collected using the API. These API helps us to gather the data for the input. Basically, it is an interface between the user and the source website from where the input tweets data could make fetched. As it's far a prolonged process, for this research purpose the data has been collected of various websites rather than collecting tweets from the Twitter itself**.**

3.2 Pre-processing

Those pre-processing of the information may be a significant step Likewise; it chooses the effectiveness of the different steps down in line. It involves syntactical correction of the tweets as desired.

Data obtained from twitter is not fit for extracting features. Mostly tweets consists of message along with usernames, empty spaces, special characters, stop words, emoticons, abbreviations, hash tags, time stamps, URL's ,etc. Thus to make this data fit for mining.

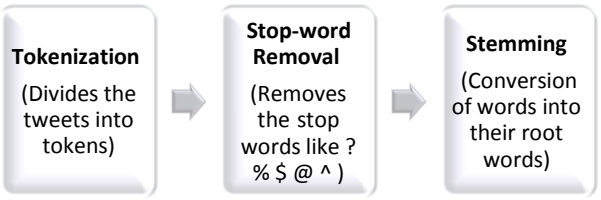

Fig (2): Data pre-processing Steps

3.3 Feature Extraction

Various methodologies for extracting features are available in the present day. The pre-processed data set need a number of dissimilar properties. In the feature extraction strategy, we extricate those parts starting with the processed data set. It identifies the features of the enter object. It is possible that any object is data in shape of text, image or video. Vector space model is mostly used as text data model. It represents the textual content into the form of vectors. It gives an independent dimension to each word.

Clustering is a process in which features that bring comparable properties are divided into the small segments or groups. In feature extraction process, assign the labels to each of the features in the cluster by using TF-IDF (Term Frequency- Inverse Document Frequency) is a numerical statistic that reflects the value of a word for the whole document (here, tweet) and is a totally efficient approach and is extensively utilized in textual content classification and data mining.

Proposed a method they don't simply depend upon vocabulary used however additionally the expressions and sentence shape utilized in unique conditions. TF-IDF measures the frequency of the term in the whole document. Here, the end results selection of useful words from tweets is feature extraction, are the labeled features for the optimization process.

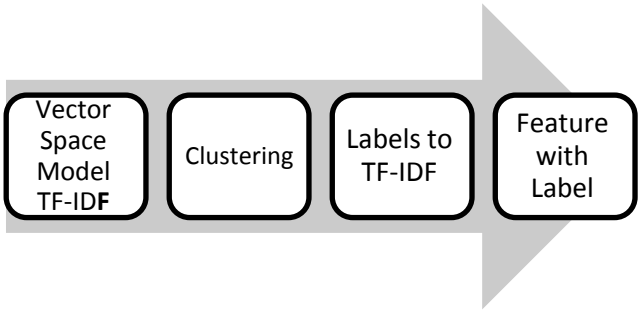

Fig (3) : Feature Extraction process

3.4 Optimization

In this paper, ACO and PSO are utilized to for the optimization process. This phenomenon is referred to as 'combinatory optimization'. It may be multi-objective functions that need which means of searching on the preliminary values aiming to decrease the final results of our function. It is a process in which the relevant features are selected from the set of the feature. This process is done by using some algorithm which is done better in optimization Ant colony optimization and Particle swarm optimization both are based totally at the biological behavior of the ants and swarms. By this technique it found the most brief way or routes of ants. Those yield for the experiment demonstrate that the optimize algorithm not just reduce the number for paths in the ACO, but also discovering the briefest way at the place of largest path. These algorithms use the same principle that are utilized by the ants and provides us optimized features

3.5 Supervised Classifiers

In this stage, classify the features as stated by their properties. Classification in this work is done by way of the usage of the Support Vector Machine and Naïve Bayes Classifier.

- **Naïve Bayes***:*
- The Naïve Bayes classifier in a standout amongst the simplest probabilistic model meets expectations positively on text classification and utilized looking into Bayes rule with self-supporting feature collection [13] meets expectations positively around quick text classification and works on Bayes rule.

With Self-supporting feature collection [13], it is flexible in way of handling with whatever number of classes or attributes. For a given tweet d, C* is a class variable which defines the sentiment given by

$$C^* = \text{argMax}_C P_{\text{NB}}(\text{C}|\text{D}) \quad (4.1)$$

Bayes Probability $P_{\text{NB}}(\text{C}|\text{D})$ described as

$$P_{NB}(\text{C}|\text{D}) = \frac{(P(C)\sum_{i=1}^{m} P(F\backslash C)^{n_i(d)})}{P(d)} \quad (4.2)$$

Here, f is feature and $n_i(d)$ is feature count found in d, m represents total number of features and P(c) and P(f|c) are found through maximum likelihood estimates [14].

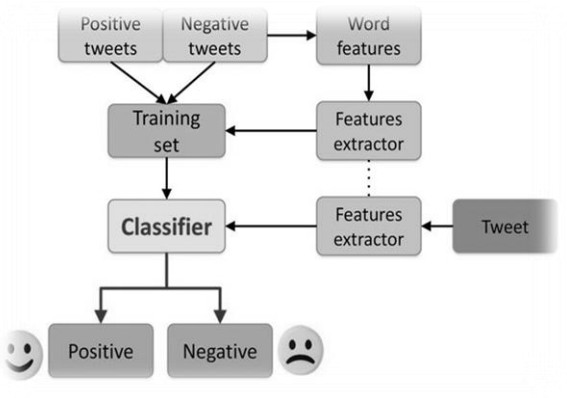

Fig. (4): Flowchart for Supervised Classifiers

- During classification phase we found a word which was not found in training phase then we will give zero as probability for positive, negative and neutral classes. To end this problem, we tend to make probability equal using Laplacian smoothing constant *k*=1.

$$\frac{Term\_count + k}{Tota\_Terms + k|c|} \qquad (4.3)$$

- ▶ **Support vector machines:**

- Support vector machines (SVM) is a blend of a linear modeling Furthermore occurrence-based learning in a high-dimensional space. SVM may be carried out for the ones problems whilst data can't make separated by way of line. SVM use nonlinear mapping – It transforms the instance space into some another space which need higher size over the first.

Kernel idea gave upward push to support vector machines. Kernel is a function which fulfil mapping of a nonlinear data to another space.

Kernel function *K* will be an inward item Φ(*x*) • Φ(*y*) between of two points *x* and *y*:

$$K(x,\ y)\ =\ \Phi(x)\ \bullet\ \Phi(y) \qquad (4.4)$$

where Φ (*x*) and Φ (*y*) are mapping operators.

The feature, that kernel characteristic is formulated as an internal product, offers a possibility to update scalar product with a few preferences of kernel [15].

The problem of finding parameters of SVM steady with a convex optimization trouble, which means that that nearby result may be is global optimum as well.

In general, the categorization task usually dividing data under traineeship and experiment sets.

Those objective of SVM may be to prepare a model (primarily based at the traineeship data).

SVM for classification may be utilized to discover a linear model of the following form:

$$y(x)\ =\ w^{T}x\ +\ b \qquad (4.5)$$

wherein x is enter vector, w and b are parameters which may be modified for a certain model and estimated in an experimental method. In simple linear classification type may be to cut down a regularized error function given by means of Equation 4.6.

$$C\sum_{n=1}^{N}\varepsilon_n + \frac{1}{2}\|\omega\|^2 \qquad (4.6)$$

whereas $\xi_n \geq 0$, $\forall\ n = 1, \ldots, N$, and

$$y(\omega^{\mathrm{T}}\mathrm{x} + \mathrm{b}) \geq 1 - \varepsilon_n \qquad (4.7)$$

Figure 3.5 illustrates a sample of a linear SVM has been trained on examples from two classes. Here the SVM constructs an isolating hyper plane and then tries to maximize the" margin" betwixt the two classes. With figure that margin, the SVM constructs two parallel hyper planes, one on each side of the initial one. These hyper planes are then "pushed" perpendicularly away from one another until they come in contact with the closest examples from either class. These examples are referred to as SVM and are illustrated in bold in Figure 3.5.

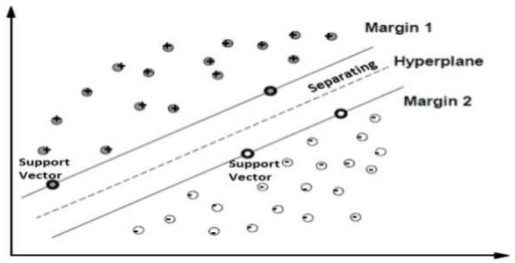

Fig. (5): Support Vector Machine: Classification

- ▶ **Kernel Functions:**

Following are forms of SVM kernel functions used for the categorization. In research [16] the four following basic kernels are described:

- Linear kernel:
$$K(x,y) = x^{T}Y + C \qquad (4.8)$$
- Polynomial kernel :
$$K(x,y) = (x^{T}Y + C)^{d} \qquad (4.9)$$
- Radial basis kernel :
$$K(x,y) = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \qquad (4.10)$$
- Sigmoid kernel :
$$K(x_i,y) = \tanh(\gamma x^{T}y + C) \qquad (4.11)$$

3.6 Performance Evaluation:

Should figure that accuracy to classifier, we required to measure which accuracy might a chance to be acquired. There are two measures on which accuracy may be dependent: **Accuracy, Precision, Recall, F Score.**

Prediction Class \ True Class	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Table (1): Confusion Matrix

The result is produced from where of Precision, Recall, F score, and Accuracy.

- **Precision**: It is the proportion of documents of rightly classified under positive prediction class to all documents under positive prediction class.

$$Precision = \frac{TP}{TP+FP} \qquad (4.12)$$

- **Recall:** It is the proportion of documents of rightly classified under positive prediction class to the documents that are positive in the negative prediction class.

$$Recall = \frac{TP}{TP+FN} \qquad (4.13)$$

- **Accuracy:** Accuracy and precision are two vital variables important factors to consider when bringing data estimations, we have to discover the accuracy of classifiers. Accuracy for any prediction model can be given as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4.14)$$

- **F-Score**: (**also F-score or F-measure**) is the weighted common of Precision and Recall. Therefore, F-Score takes both false positives and false negatives into consideration. "Fscore isn't always as easy to apprehend as accuracy, however it's far lot extra beneficial than accuracy, especially Assuming that you have got a uneven class distribution. Accuracy works best in cases like if the false positives and the false negatives have comparable cost.

$$F\ SCORE = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \qquad (4.15)$$

4. RESULTS AND DISCUSSIONS

This chapter displays and analysis the test outcomes and the assessment for our approach. First, we compare results of different methods applied for the sentiment analysis of data obtained from Twitter. Second, the discussions on the effects of various features are presented. Also, we discussed the best obtained results (which were given by SVM and NB methods). Finally, we proposed further work as there is at present a considerable measure of room for improvement.

- **Comparison of NB, NB-ACO, NB-PSO results in the form of bar chart having x-axis containing precision, recall, accuracy and y-axis contains percentage:**

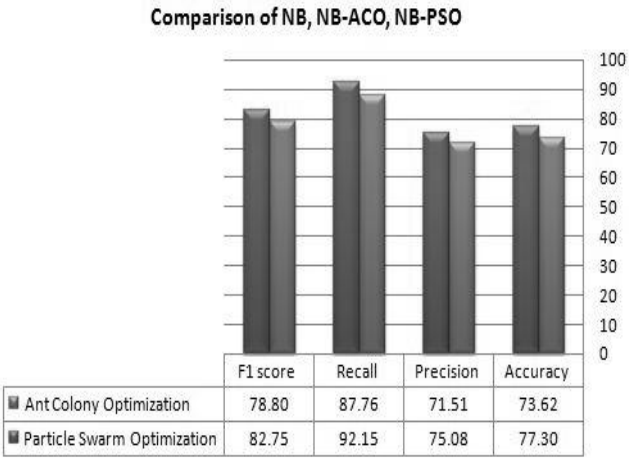

Fig. (6): Graph of results NB_ACO versus NB_PSO

- **Comparison of SVM, SVM-ACO, SVM- PSO results in the form of bar chart having x-axis containing precision, recall, accuracy and y-axis contains percentage:**

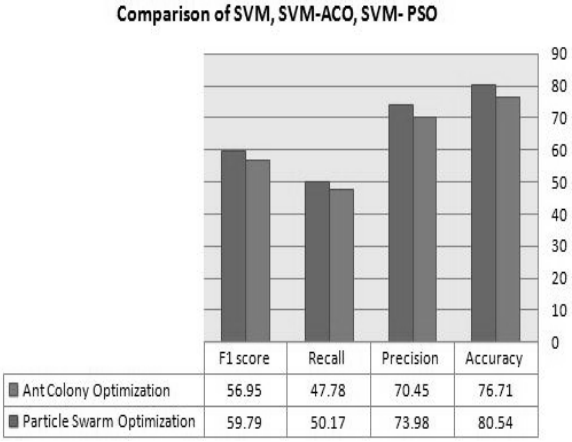

Fig. (7): Graph of results SVM_ACO versus SVM_PSO

5. CONCLUSION

Sentiment analysis is used to identify people's opinion, attitude and emotional states. The views of the human beings (man/woman) may be positive or negative. Commonly, elements of speech are used as function to extract the sentiment of the textual content. An adjective performs a essential function in figuring out sentiment from components of speech. Sometimes words having adjective and adverb are used together then it is difficult to identify sentiment and opinion.

I established, examined and evaluated sundry machine learning methods for the Sentiment Analysis task. I learned a considerable measure for things about how to face a machine learning trouble and how to do data analysis to make the work easier to the machine to learn.

I see that of the almost important things when we are facing a text classification problem is the type of text and the phrases that we will see in the data. This is because it has an important impact on the wide variety of words that the machine learning

technique will learn and, in consequence, the final number of features.

In this document, analysis is done by the effective weight by PSO and ACO with discriminative classifier of SVM and Naïve Bayes. Results shows SVM_PSO perform but not well comparison of Naive Bayes which is 77.30% accuracy because Naive Bayes iteratively change the threshold if weight is different for different keywords.

The accuracy level can nonetheless be improved by considering the emoticons for the categorization of the input data as it can help a lot for correct category classification and also by using another optimization technique along with the classifiers.

For future work, we would like to make bigger the domain of our experiments and run the classifiers on multiple dataset considering number of different languages so as will have more representative inputs and thus better generalizable results.

6. REFERENCES

[1] A Gupta, J Pruthi, N Sahu" Sentiment Analysis of Tweets using Machine Learning Approach " International Journal of Computer Science and Mobile Computing, Vol.6 Issue.4, April- 2017, pg. 444-458.

[2] B Naiknaware, B Kushwaha, S Kawathekar , "Social Media Sentiment Analysis using Machine Learning Classifiers" International Journal of Computer Science and Mobile Computing, Vol.6 Issue.6, June- 2017, pg. 465-472.

[3] Andrew, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods20016Nello Christianini and John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000. xiii + 189 pp., ISBN: ISBN 0-521-78019-5 Hardback:£27.50", Kybernetes, vol. 30, no. 1, pp. 103-115, 2001.

[4] M. Anjaria and R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning", Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on, pp. 1--8, 2014.

[5] J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis", Proceedings of the Third International ICWSM Conference, vol. 9, 2009.

[6] SM. Mohammad, S. Kiritchenko and X. Zhu, "NRC-Canada: Building the State-of- the-Art in Sentiment Analysis of Tweets", arXiv preprint arXiv:1308.6242, 2013.

[7] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By," Sentiment Analysis on Social Media ", Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, pp.919—926, 2012.

[8] P Gamallo, M Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[9] P Liang, B Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN:978-1-494673-6068-5, http://doi.ieeecomputersociety.org/10.1109/MDM.2013.

[10] 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.

[11] G Beigi1 Analysis in Social Media and its Applications in Disaster Relief, Computer Science and Engineering, Arizona State University , Texas A&M University 2{hu}@cse.tamu.edu,p.2-3

[12] H. Wang, D. Can, A. Kazemzadeh, F. Bar and S. Narayanan," A System for Real- time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle",Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics, pp. 115—120, 2012.

[13] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou and P. Li, "User-level sentiment analysis incorporating social networks", Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1397--1405, 2011.

[14] M. Koppel and J. Schler, "THE IMPORTANCE OF NEUTRAL EXAMPLES FOR LEARNING SENTIMENT", Computational Intell, vol. 22, no. 2, pp. 100-109, 2006.

[15] O Chapelle. Support vector machines et classification d'images. 1998.

[16] C Wei Hsu, C Chung Chang, C Jen Lin, et al. A practical guide to support vector classification. 2003