Twitter Sentiment Analysis using Machine Learning and Optimization Techniques

Prachi Bansal
M.Tech Scholar, CSE Department
Doaba Institute of Engg. & Technology, Kharar
Punjab, India

Ramanjot Kaur
Assistant Professor, CSE Department
Doaba Institute of Engg. & Technology, Kharar
Punjab, India

ABSTRACT
Sentiment Analysis means determining the views of the user from the text regarding that topic i.e. how one feels about it. It can be used to classify the text content into positive or negative. Various researchers have used a wide range of methods to train the classifiers for the Twitter dataset with varying results. This paper introduces a hybrid approach 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 Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) before performing the binary text categorization. Naïve Bayes and Support Vector Machine (SVM) is the machine learning techniques used for the binary classification of tweets. The results drawn using optimization with classifiers is much efficient than using classifier alone.

Keywords
Sentiment Analysis, Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Naïve Bayes, SVM.

1. INTRODUCTION
In the era of advanced Internet Technology, people from all over the world connect to each other through social networking sites like Facebook, Orkut, Twitter, etc. Twitter has nearly 300 million active users with Feature Optimization has been performed using 500 million tweets per day; which makes it leading social networking website worldwide. As Twitter has this huge number of users and data, it has always been used as an informative resource by various organizations to research public opinions and gather critical feedback [2].

In a tweet, users can write their views regarding any topic or general thoughts in a maximum length of 140 characters only. Due to this limited tweet length, people write in a very concise manner by using slangs; which makes sentiment analysis a tedious task.

Sentiment Analysis can be defined as the process of categorizing the opinions expressed through tweet to understand the user views about that topic. Their opinions can be positive, negative or neutral. It is beneficial for the marketers as they can analyse the opinion of the public towards their brand and existing/newly released products; which would help them to evaluate their performance and improve it [3]. A lot of research has already been done in the sentiment analysis on Twitter by using various machine learning methods. But the required results have still not been achieved.

Sentiment Analysis can be performed in a series of steps which are Data Collection, Pre-processing, Feature Extraction, Feature Optimization, and Classification. Swarm Intelligence Optimization Techniques are used which are inspired by the behaviour of the group of insects in our nature. Ant Colony Optimization (ACO) techniques inhibits the natural behaviour of ants which aims to find the food through the shortest 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 has been followed very frequently. The same behaviour has been incorporated in the computer artificial ants to search the good solutions for a given problem.

Particle Swarm Optimization (PSO) inhibits the social behaviour of birds as they intend to live in flocks. All the birds try to find the food in an area. But they know how far the food is. So they likely to follow the bird who is nearest to the food. This behaviour has been incorporated to improve the candidate solution iteratively locally and globally which helps to find out the best optimized solution.

Machine Learning methods are used to classify the tweets into positive or negative categories. The best results are given by using the supervised machine learning model. Naïve Bayes and Support Vector Machine are two learning methods considered for this research. Naïve Bayes is a classifier based on the Bayes Theorem to calculate the probability of positive and negative category for a given tweet and outputs the category having the higher value. SVM is based on statistical learning model which aims to divide the hyper-plane into two classes.

To perform sentiment analysis on Twitter, the datasets are collected from various websites like Stanford University. Then the collected data has been pre-processed to make it more understandable. Pre-processing includes conversion of tweets into lowercase; tokenization which is splitting of sentences into words; removal of stop words and emoticons; stemming which is replacing words into their root words. After the labelled data has been pre-processed, the Term Frequency-Inverse Term Frequency(TF-IDF) for each word has been calculated and made a feature vector for each tweet. This feature vector can be optimized with the help of ACO and PSO. This feature vector has been fed to the classifiers to learn the process of binary-class (positive and negative) text categorization.

The model has been implemented 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 concerned with the identification and

classification of sentiments. Social media is the platform which is used to express the feeling related to any entity. Knowledge-based approach and machine learning approach is used for the sentiment classification [1]. Sentic computing is knowledge-based computing which ensemble the application of common sense computing. This concept is also known as concept level sentiment analysis. For improving the accuracy of the task like polarity detection the machine learning concept is used [2]. SentBuk is Facebook applications which are used for retrieval of the messages written by the users and classify them according to the polarity. It is a hybrid approach which combines the lexical-based and machine learning techniques [3]. Detection of sentiment is performed at tweet level and entity level both. Lexicon based approach is used for the sentiment analysis of Twitter data which provides a fixed and static polarity to the tweets. The representation of these words is called Senti circle [4]. NLP (Natural Language Processing) approach is used to identify the fine-grained application features in the reviews. Sentiments are extracted from the reviews in the form of features and a general score value is given to all the features. This method helps to analyse the relevant reviews with better accuracy rate [5]. An ontology-based method is used for the sentiment analysis which is the more efficient method. The proposed method not only characterized by the sentiment score but it also provides sentiment grade for each post. This method provides a detailed analysis of opinion on the specific topic [6].

CNN model proposed tree bank information for the sentiment analysis. It does not consider the syntax information but it works on the structural information which performs better than single factor model. Deep convolution neural network worked with character level sentiment analysis which provides an effective polarity of tweets in different languages. This model is able to learn the latent features from all languages [7]. CNN is used with multiple filters with varying window size on the top of the connected layers. It works on the unsupervised learning method. This approach does not depend on the manually selected features [9]. Supervised learning classifiers are mainly used for classification of polarity on feature vector which is extracted from the text. Text preprocessing is the first process which is performed in the sentiment analysis. After that feature selection and classification is done by the SVM (support vector machine) classifier [10]. Reviews posted by the consumers are very important for the decision making in business perspective. Gini index is a method of feature selection which is used to select the features from the text posted by the consumer's post. Support vector machine classifier is used to classify the features. SVM with Gini Index method provides the high accuracy rate with low error rate.

3. METHODOLOGY

The methodology followed in this research has been explained in various steps below

3.1 Data Collection

It is a process in which the input data can be collected from the source by using the API. These API helps us to collect the data for the input. Basically it is an interface between the user and the source website from where the input tweets data can be fetched. As it is a lengthy process, for this research purpose the data has been collected from various websites rather than collecting tweets from the Twitter itself.

3.2 Pre-processing

Data pre-processing is also called as Data cleaning. In this process, noisy, incomplete and inconsistent data is removed. For this following process are used:

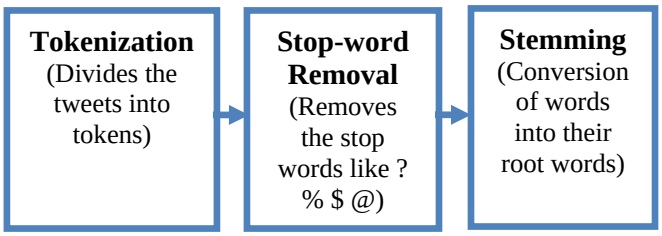

Fig.1: Pre-processing Steps

3.3 Feature Extraction

It is a process in which raw data is transformed into the features. It identifies the features of the input object. An object may be any data in form of text, image or video. Vector space model is mostly used as text data model. It represents the text into the form of vectors. It gives an independent dimension to each word.

Clustering is a process in which features that have similar 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). TF-IDF measures the frequency of the term in the entire document. Here, the end results are the labelled features for the optimization process.

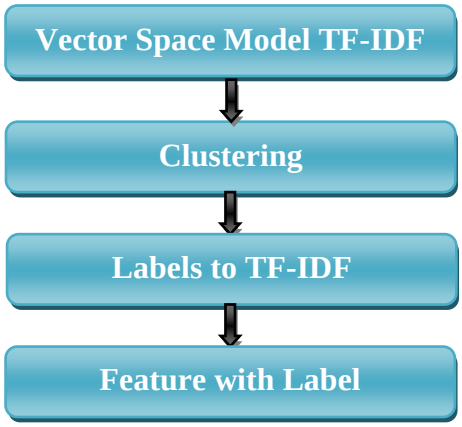

Fig.2: Feature Extraction Process

3.4 Optimization

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 performs better in optimization. In this paper, ACO and PSO are used for the optimization process. Ant colony optimization and Particle swarm optimization both are based on the biological behaviour of the ants and swarms. These algorithms use the same principle that are used by the ants and provides us optimized features.

3.5 Classification

In this stage, classify the features according to their properties. Classification in this work is done by using the Support Vector Machine and Naïve Bayes Classifier.

Naïve Bayes classifier is a classification algorithm which is based on the Bayesian theorem. It is used to predict the probability of each feature belongs to the cluster.

Computing probability for each class by using equation 1.

$$p(b_y|a_n^d) = \frac{p(b_y)p(a_n^d|b_y)}{\sum_{i=1}^{c} p(b_x)p(x_n^d|b_x)}, \text{ y=1,2……..c} \quad (1)$$

Where
$p(b_x)$ is the y_i prior probability,
$p(x_n^d | b_y)$ ←Conditional class probability density function.

SVM classifier is a binary classifier it transforms the data to the higher dimension by using kernel function. It constructs the hyper-plane to form the decision boundary. It uses the labelled features for classification. The data points which define the boundary are called as support vectors.
Following are types of SVM kernel function used for the classification.

Polynomial Kernel: $K(y_{i,}, y_j) = K(y_{i,}, y_j + 1)^h$

Sigmoid Kernel: $K(y_{i,}, y_j) = tanh(ky_{i,}, y_j - \delta)^h$

RBF Kernel: $K(y_{i,}, y_j) = e^{-||y_i - y_j||^2/2\sigma^2}$

Here $K(y_{i,}, y_j)$ *is kernel function*

ALGORITHM

Step 1: Input the tweet text.
Step 2: Stream the Tweet text.
Step 3: Pre-process the input data.
$Tokenization \rightarrow Stop\ word\ removal \rightarrow Stemming$
Step 4: Build the vector space model TFIDF

$$tf(t,d) = \frac{f_d(t)}{\min_{\omega \epsilon d} f_d(\omega)}$$

$$idf(t,d) = ln\left(\frac{|D|}{|\{d \epsilon D : t \epsilon d\}|}\right)$$

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

$f_d(t) :=$ *frequency of term t in document d*
$D :=$ *Corpus of documents*

Step 5: Make the cluster of the features.
Step 6: Label the TF-IDF to the features.
Step 7: Use the meta-heuristic ACO and PSO for the optimization of the features.
Step 8: Classifies the optimized features using Support Vector Machine and Naïve Bayes Classifier.
Step 9: Test the classifier model.
Step 10: Analyse the Accuracy, Precision and Recall.

4. RESULTS AND DISCUSSION

4.1 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

Table 1. Result of NB_ACO

Parameters	Value
Accuracy	85.05
Precision	85.60
Recall	84.51
F Score	85.05

Table 2. Result of NB_PSO

Parameters	Value
Accuracy	86.29
Precision	86.80
Recall	82.60
F Score	86.31

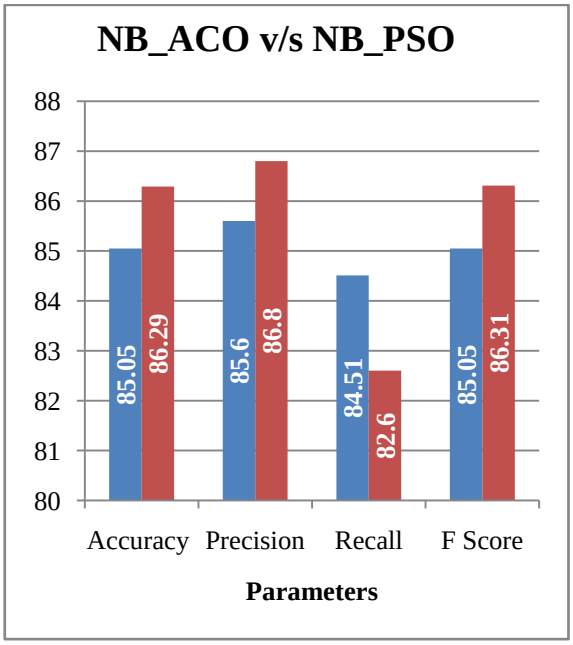

Fig.3: Graph of results NB_ACO versus NB_PSO

4.2 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.

Table 3. Result of SVM_ACO

Parameters	Value
Accuracy	82.35
Precision	89.67
Recall	86.01
F Score	87.80

Table 4. Result of SVM_PSO

Parameters	Value
Accuracy	84.21
Precision	81.3
Recall	80.00
F Score	80.64

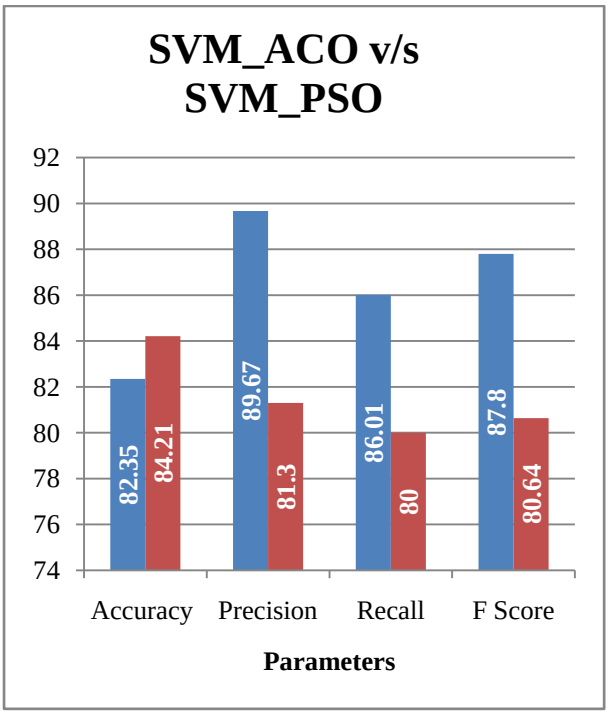

Fig.4: Graph of results SVM_ACO versus SVM_PSO

5. CONCLUSION

Sentiment analysis is a challenging problem because of automatic identification of text semantic, so it is essential to learn the pattern of text and give the effective weight to every keyword. In this paper, analysis is done by the effective weight by Particle Swarm Optimization and Ant Colony Optimization with discriminative classifier of SVM and Naïve Bayes. Results shows SVM_PSO perform but not well comparison of Naive Bayes which is 86.29% accuracy because Naive Bayes iteratively change the threshold if weight is different for different keywords. The accuracy level can still be improved by considering the emoticons for the classification of the input data as it can help a lot for correct category classification and also by using other optimization methods along with the classifiers.

6. REFERENCES

[1] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.

[2] Poria, Soujanya, et al. "Sentic patterns: Dependency-based rules for concept-level sentiment analysis." Knowledge-Based Systems 69 (2014): 45-63.

[3] Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." Computers in Human Behaviour 31 (2014): 527-541.

[4] Saif, Hassan, et al. "Contextual semantics for sentiment analysis of Twitter." Information Processing & Management52.1 (2016): 5-19.

[5] Guzman, Emitza, and Walid Maalej. "How do users like this feature? a fine grained sentiment analysis of app reviews." Requirements Engineering Conference (RE), 2014 IEEE 22nd International. IEEE, 2014.

[6] Kontopoulos, Efstratios, et al. "Ontology-based sentiment analysis of twitter posts." Expert systems with applications40.10 (2013): 4065-4074.

[7] Yang, Tao, et al. "Tb-CNN: joint tree-bank information for sentiment analysis using CNN." Control Conference (CCC), 2016 35th Chinese. IEEE, 2016.

[8] Wehrmann, Joonatas, et al. "A character-based convolutional neural network for language-agnostic Twitter sentiment analysis." Neural Networks (IJCNN), 2017 International Joint Conference on. IEEE, 2017.

[9] Stojanovski, Dario, et al. "Twitter sentiment analysis using deep convolutional neural network." International Conference on Hybrid Artificial Intelligence Systems. Springer, Cham, 2015.

[10] Chikersal, Prerna, et al. "Modelling public sentiment in Twitter: using linguistic patterns to enhance supervised learning." International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Cham, 2015.