DOI: https://doi.org/10.24113/ijoscience.v6i11.325

Optimization of Feature Set for Sentiment Analysis using Artificial Butterfly Algorithm and Ensemble Machine Learning

Jyoti Hanvat
M.Tech. Scholar
Department of CSE
VITS
Bhopal (M.P), India
jyotihanvat99@gmail.com

Sumit Sharma
Professor
Department of CSE
VITS
Bhopal (M.P), India
sumit_sharma782022@yahoo.co.in

***Abstract*: The current decade has witnessed the remarkable developments in the field of artificial intelligence, and the revolution of deep learning has transformed the whole artificial intelligence industry. Eventually, deep learning techniques have become essential components of any model in today's computational world. Nevertheless, ensemble learning techniques promise a high degree of automation with generalized rule extraction for both text and sentiment classification tasks. This paper aims designed and implemented optimized feature matrix using ensemble learning used for sentiment classification and its applications.**

Keywords: Deep neural networks, Aspect based sentiment analysis, Accuracy.

I. INTRODUCTION

The explosive growth of opinion content generated through commercial websites and recent advances in data analytics together have placed new challenges and opportunities [1]. Investigating peculiar and potentially useful patterns from a large collection of user-generated content (UGC) is crucial for many sentiment analysis tasks [2]. Sentiment analysis techniques are specially used to recognize and extract subjective content in source data to assist an enterprise in understanding the social sentiment of their brand, product, or services [3].

However, identification of sentiments in UGC faces numerous challenges, as they are composed of incomplete, noisy, and unstructured sentences, unusual expressions, ungrammatical phrases, and non-lexical terms [4]. Besides, it is hard to explore the correlation among opinion sentences due to the diversity of linguistic issues and makes the process of sentiment analysis still more challenging [5]. Hence, to address these challenges, real-time sentiment analysis systems need to be developed for processing large volumetric opinion data in a reasonable amount of time.

Sentiment analysis methods can be generally divided into two categories, machine learning and lexicon-based methods. The former uses machine learning techniques for sentiment polarity classification. These kinds of methods usually need a lot of labeled training data. However, collecting sufficient labeled data is a challenge in itself. Lexicon based methods utilize sentiment lexicons to compute sentiment scores of given reviews. Then they group the scored reviews into positive or negative categories by the sentiment scores. Many researcher illustrated that lexicon based method is the method for constructing sentiment lexicon. Its generation is divided into two main steps i.e. dictionary-based and corpus-based approaches.

II. RELATED WORK

Yang et al. [1] proposed a new mood analysis model-SLCABG, which is based on the mood lexicon and combines the convolutional neural network (CNN) and the attention-based bidirectional recurrent unit (BiGRU). In terms of methodology, the SLCABG model combines the benefits of the sentiment dictionary and deep learning technology and overcomes the shortcomings of the existing sentiment analysis model for product reviews.

Xu et al. [2] proposed method for sentiment analysis for Big Data. This method integrates the subject's semantic information into text visualization via a neural network model. The attention mechanism is introduced into the neural network and a context-sensitive vector is introduced to calculate the weight of each word. Also, to make the model more adaptable, the mood dictionary markup method is used to get the training data.

Meyyappan et al. [3] recommend an innovative method of common sense-based sentiment analysis (domain specific ontology) for ConceptNet-based tourism ontology in Oman.

Xu et al. [4] proposed an improved word representation method that integrates the contribution of sentiment information into the traditional TF-IDF algorithm and generates weighted word vectors. Weighted word vectors are placed into long-term short-term bidirectional memory (BiLSTM) to efficiently capture context information, and comment vectors are better represented. The mood trend of

the comment is obtained from an anticipatory classifier for neural networks. Under the same conditions, the sentiment analysis proposed is compared with the sentiment analysis methods of RNN, CNN, LSTM and NB. Experimental results show that the proposed sentiment analysis method has higher accuracy, better recall and a higher F1 score. The method proved effective for very specific comments.

Iqbal et al. [5] proposed an integrated framework that bridges the gap between vocabulary-based approaches and machine learning to achieve better accuracy and scalability. To address the scalability problem that arises as the feature set grows, a new feature reduction technique based on genetic algorithms (GA) is proposed.

Wongkar et al. [6] devised a framework for analyzing Twitter data that was carried out for presidential candidates of the Republic of Indonesia in 2019. The naive Bayesian method is used to rank the classes or emotional level of society and get a score of 75.58% accuracy.

The existing methods have some deficiencies:

1. These methods only apply to some specific domains in which emoticons are used frequently
2. They need human-annotated data
3. The generated sentiment lexicon contains more positive or negative words.

In this paper, a domain-specific sentiment is proposed lexicon generation method and a sentiment analysis framework based on the generated domain-specific sentiment lexicon. The ultimate goal of sentiment analysis can be generally summarized as identifying sentiment or opinion labels of given texts. Depending on the types of final label, problems are usually divided into sentiment classification and emotion/subjectivity identification.

III. Proposed Methodology

This section explains the methodology for sentiment analysis of textual data. Overall proposed algorithm is shown in figure 1 which include three phases i.e. Dataset Preparation, Feature Extraction and Classification.

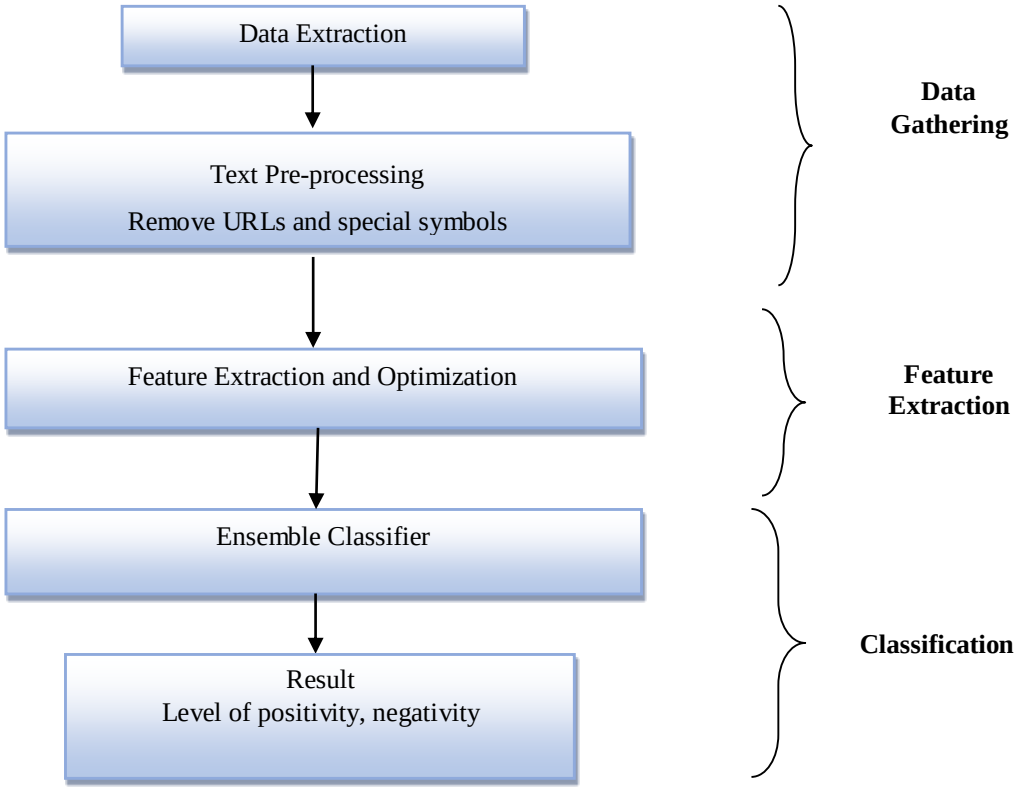

Fig.1. Proposed Flow Diagram

A. Dataset Preparation

In this section of the research methodology, a dataset is created. For data preparation, data is first collected / logged from online sources or datasets, then the dataset is preprocessed to be useful for other processes. So basically it contains:

Step 1: Data Gathering

Step 2: Text Preprocessing

The dataset must then be cleaned. This process is called data preprocessing. This step is necessary to remove unnecessary terms used by a person in alerts such as comma, period, colon, or symbols or special characters. These terms do not connect mood values. To further reduce complexity, these terms should be eliminated. In this research paper, this step is performed in two steps, as shown below:

- Sometimes url is also mentioned in reviews as suggested by any reviewer. As it is known that any url represents address, they don't convey any emotional or sentimental values.
- Some special symbols or characters are removed.
- Comma, full stops, colons and semi-colons are removed.

B. Feature Extraction and optimization

Extracting useful information from these controls is called feature extraction. This information is a very useful set of functions for classifiers. Figure 1 shows the feature extraction process for the proposed methodology. This study extracts the optimized functions, which are explained below:

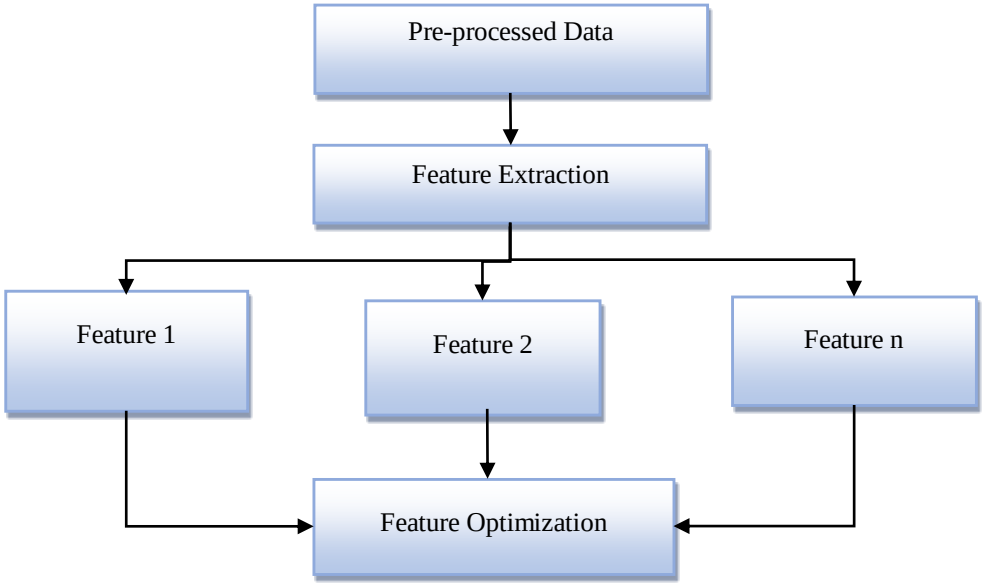

Fig.2. Feature Extraction for Proposed Work

For generation of feature vector associated with positive score and negative score of the data, sentiment data dictionary is referred. In this work sentiment score is calculated by summarizing the positive and negative score of each word in entire sentence.

Inspired on the mate-finding strategy of speckled woods, Artificial Butterfly Optimization was developed. The speckled woods prefer to live on the borders of woodlands where the sun shines on trees and create lots of sunspot. The butterfly population is sorted and divided into two groups according to their fitness. Butterflies with better fitness form the sunspot butterflies and the rest form canopy butterflies, and a different flight strategy is applied to each group.

Two modes compose the ABO algorithm:

Sunspot mode

Canopy mode

Some rules of butterflies in ABO algorithm are stated as below:

- In order to increase the likeliness of encountering female butterflies, all male butterflies attempt to fly toward a better location called a sunspot
- To occupy a better sunspot, each sunspot butterfly always attempts to fly to its neighbor's sunspot.
- Each canopy butterfly continually flies toward any sunspot butterfly to contend for the sunspot.

Let P = $\{p_1, p_2, \ldots, p_m\}$ = Population of butterflies.

The following strategy is used for the sunspot mode or the canopy mode. Each butterfly flies toward a randomly selected butterfly as follows:

$$P_i^{n+1} = (P_i^n - P_k^n)\beta \quad (1)$$

Where, i= ith butterfly

n = iteration

β = random generated number between $[1, -1]$

k = randomly selected butterfly ($k \neq i$)

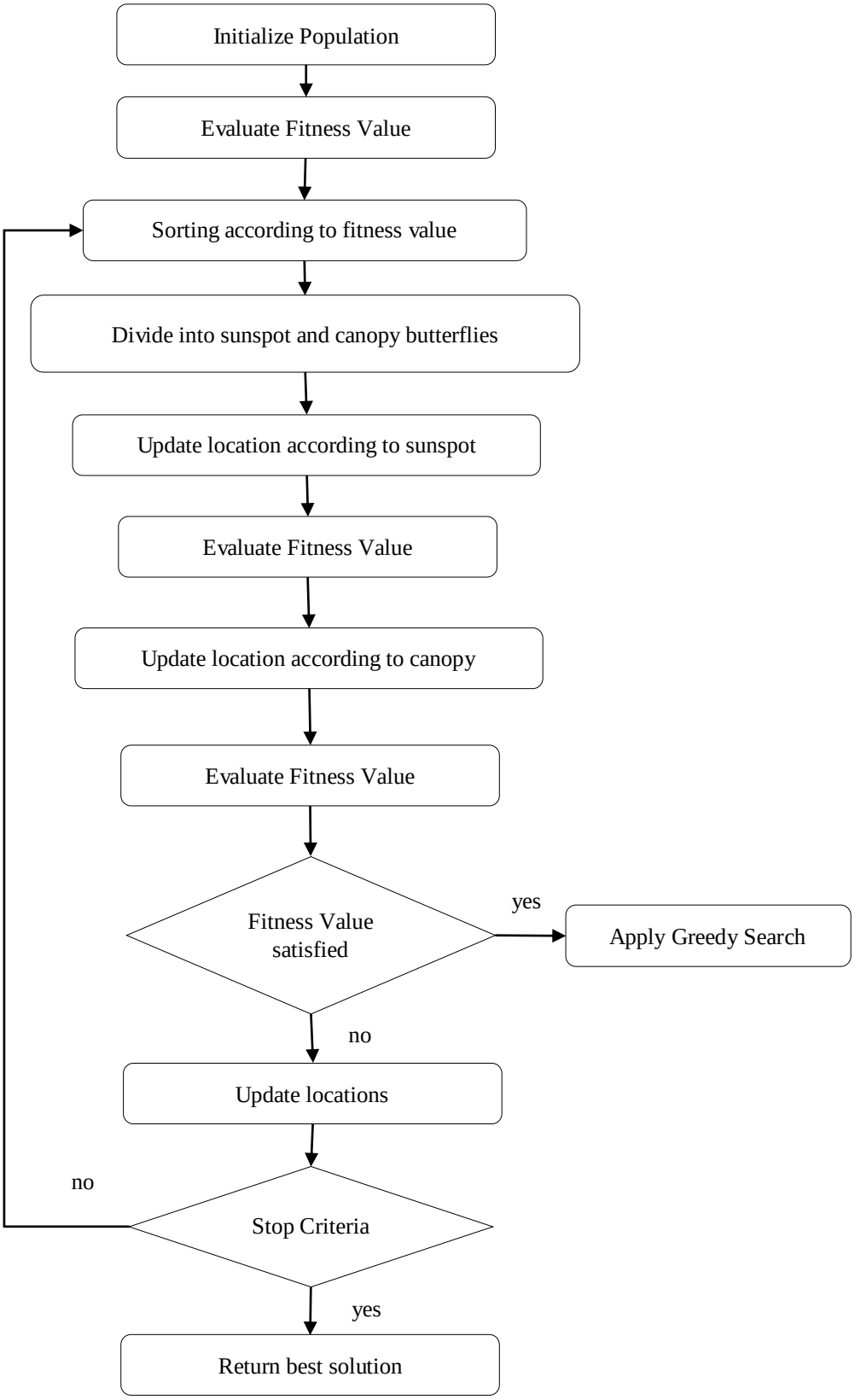

Fig.3. Algorithm for Artificial Butterfly Optimization

Each butterfly flies toward a randomly selected sunspot butterfly as follows:

$$P_i^{n+1} = P_i^n + \frac{P_k^n - P_i^n}{x_k^n - P_i^n})(Ub - Lb)s\beta \quad (2)$$

Where, Ub= upper bound

Lb= Lowe bound

The s parameter decreases linearly from 1 to s_e , as follows:

$$s = 1 = (1 - S_e)\frac{n}{N} \quad (3)$$

where N = Max iteration

C. Classification

Classification technique is used to categorize any data values into different classes. In this research work ensemble learning approach is used to classify dataset into positive or negative class.

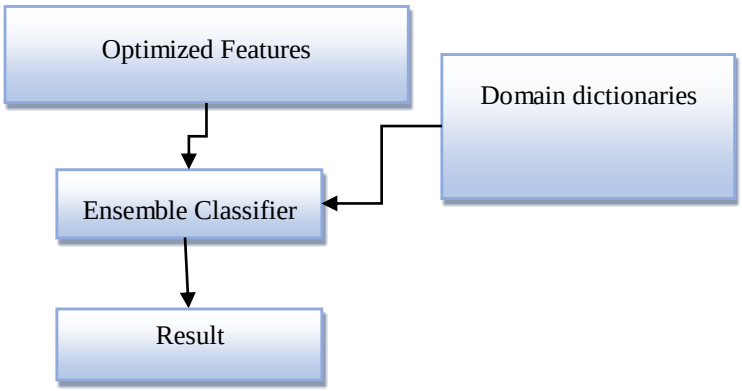

Fig.4. Feature Classification for Proposed Work

In this research work, the dataset is prepared and simulation is performed using proposed algorithm. For performance evaluation, Ensemble classifier are used which are discussed below:

Ensemble Support Vector Machine

The support vector machine (SVM) algorithm has some drawbacks. One of the problems associated with SVM is that it was designed and well suited for binary-class classification problems. For multi-class classification, SVM has to be combined multiple times. After combining, SVM multiple times, its performance degrades and as per study ensemble techniques gives more efficient result as compared to traditional SVM method. Another drawback of SVM is that during learning process with large feature sets containing dataset, it consumes much processing time to converge. In order to reduce such time complexity approximation algorithms are needed to be applied. The use of any approximation algorithms would degrade the performance efficiency of SVM classification. In order to overcome the drawbacks of traditional SVM machine learning algorithm, this paper proposed and ensemble SVM classification algorithm for multilevel classification of intrusion detection system as discussed in above section. The proposed ensemble SVM improves the efficiency and accuracy rate of the classification problem.

In ensemble SVM, each SVM module is trained independently with random training samples and correctly classified the data samples of each SVM. Similarly, all other data values are trained independently on individual SVM module and finally integrated as an ensemble or combination of several SVMs which will expand the correctly classified area incrementally. This proposed ensemble SVM will performs better in case of multi-class classification problems. In fig. 3, a generalized architecture of proposed SVM is given and explained.

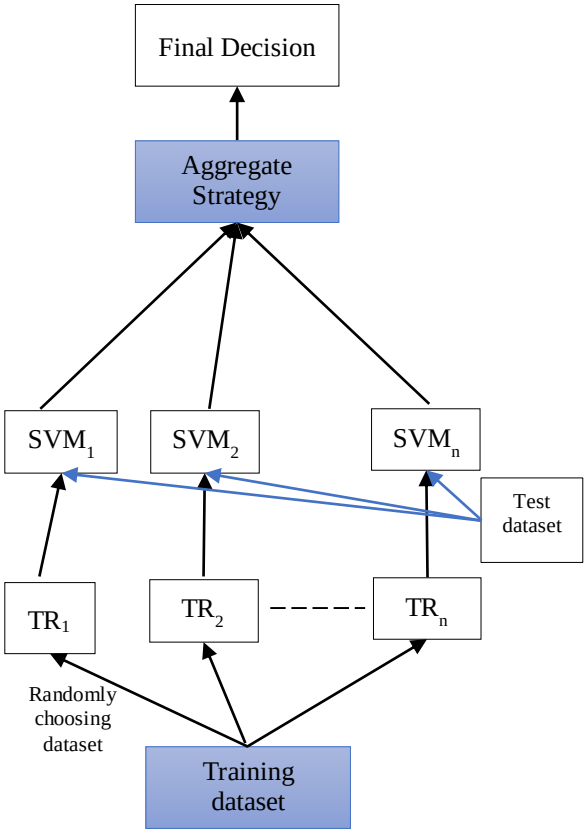

Fig.5. Architecture of the Ensemble SVM

During training phase of proposed ensemble SVM architecture, each SVM module is trained individually with randomly selected training samples from the dataset. This makes each trained SVM module be different from each other. Each SVM module can be trained with different

training sets and rules. Bagging, random selection and boosting selection strategies can be used to select training samples. In this proposed architecture bagging rules are taken as a base for ensemble SVM in which each SVM modules are trained individually and further they are aggregated by applying combination method. During testing phase, the aggregate strategy or voting strategy among all SVM module will decide the test data class label. In ensemble SVM architecture, "n" training samples sets are constructed with "n" individual SVMs modules. To achieve higher efficiency, different training sample sets are taken in order to improve the aggregation result with higher efficiency.

IV. Result Analysis

This section comprises with an analytical and numerical description of proposed algorithm for text sentiment analysis which is simulated to obtain the performance of the proposed algorithm. In order to evaluate the performance of proposed algorithm scheme, the proposed algorithm is simulated in following configuration:

Software Requirement

MATLAB-8.3.0 Platform

32/64 bit Windows Operating System

Hardware Requirement

Intel Core i5-3210M CPU @ 2.50GHz

2 GB RAM

512 GB Hard Disk

Some of the performance parameters are discussed below:

Recognition Accuracy is represented as:

$$(TP+TN)/(TP+TN+FP+FN) \quad (4)$$

Precision is represented as:

$$(TP)/(TP+FP) \quad (5)$$

Recall is represented as:

$$(TP)/(TP+FN) \quad (6)$$

F_measure is represented as:

$$(2*Recall*Precision)/(Recall + Precision) \quad (7)$$

Where,

TP= True Positive, that means if given text sample is of positive type and predicted value also shows that it is positive.

TN= True Negative, that means if given text sample is of negative type and predicted value also shows that it is negative.

FP = False Positive, that means if given text sample is of negative type and predicted value also shows that it is positive.

FN= False Negative, that means if given text sample is of positive type and predicted value also shows that it is negative.

Table 1 shows the performance evaluation of proposed algorithm with and without optimization over datasets. From the result analysis it has been analyzed that with optimization the classification achieved best result.

Table 1: Performance Evaluation of Proposed Algorithm

Algorithms	Accuracy
With Optimization	97.45
Without Optimization	85.34

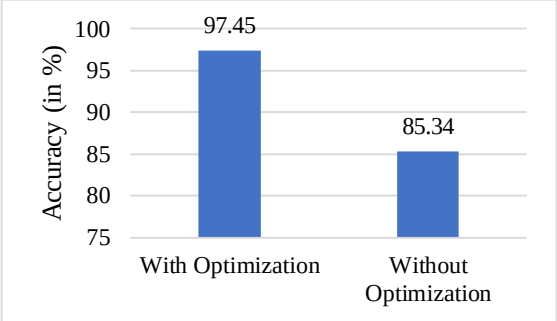

Fig.6. Performance Comparison of Accuracy

Table 2 shows the comparative performance evaluation of proposed algorithm with optimization and existing algorithm. From the result analysis it has been analyzed that proposed algorithm had achieved best result.

Table 2: Comparative Performance Evaluation

Algorithms	Accuracy
ABA	97.45
GA	95.7

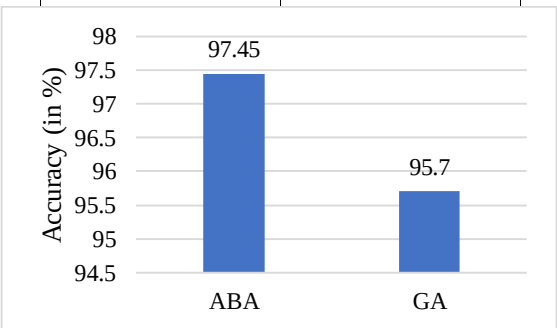

Fig.7. Performance Comparison of Accuracy

V. CONCLUSION

In this paper, optimized sentiment analysis framework is proposed for different review dataset. In this paper, to solve the scalability issue that arises as the feature-set grows, feature reduction technique is proposed using swarm intelligence, termed as artificial butterfly algorithm, are employed which includes ensemble machine learning approach with optimized feature selection. The result analysis shows improvement over existing work.

VI. REFERENCES

[1] L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in IEEE Access, vol. 8, pp. 23522-23530, 2020.

[2] G. Xu, Z. Yu, Z. Chen, X. Qiu and H. Yao, "Sensitive Information Topics-Based Sentiment Analysis Method for Big Data," in IEEE Access, vol. 7, pp. 96177-96190, 2019.

[3] V. Ramanathan and T. Meyyappan, "Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism," 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, Oman, 2019, pp. 1-5.

[4] G. Xu, Y. Meng, X. Qiu, Z. Yu and X. Wu, "Sentiment Analysis of Comment Texts Based on BiLSTM," in IEEE Access, vol. 7, pp. 51522-51532, 2019.

[5] F. Iqbal et al., "A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction," in IEEE Access, vol. 7, pp. 14637-14652, 2019.

[6] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm of The Data Crawler: Twitter," 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 2019, pp. 1-5.

[7] X. Hu, J. Tang, H. Gao, and H. Liu, ''Unsupervised sentiment analysis with emotional signals,'' in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 607–618.

[8] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, ''Sentiment analysis of Twitter data,'' in Proc. Workshop Lang. Social Media, 2011, pp. 30–38.

[9] M. Pontiki et al., ''SemEval-2016 task 5: Aspect based sentiment analysis,'' in Proc. 8th Int. Workshop Semantic Eval. (SemEval), 2014, pp. 27–35.

[10] P. C. S. Njølstad, L. S. Høysæter, W. Wei, and J. A. Gulla, ''Evaluating feature sets and classifiers for sentiment analysis of financial news,'' in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT), vol. 2, Aug. 2014, pp. 71–78

[11] S. Shikhar, et al. "LEXER: LEXicon Based Emotion AnalyzeR." International Conference on Pattern Recognition and Machine Intelligence. Springer, Cham, 2017.

[12] Low, Lu-Shih Alex, et al. "Content based clinical depression detection in adolescents." Signal Processing Conference, 2009 17th European. IEEE, 2009.

[13] Wang, Xinyu, et al. "A depression detection model based on sentiment analysis in micro-blog social network." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2013.

[14] Wang, Xinyu, Chunhong Zhang, and Li Sun. "An improved model for depression detection in micro-blog social network." 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW). IEEE, 2013.

[15] Shen, Tiancheng, "Cross Domain Depression Detection via Harvesting Social Media." Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. 2018.