



A NOVEL SENTIMENT ANALYSIS FOR AMAZON DATA WITH TSA BASED FEATURE SELECTION

ANAND JOSEPH DANIEL D.*AND JANAKI MEENA M.†

Abstract. Sentiment analysis of online product reviews has become a mainstream way for businesses on e-commerce platforms to promote their products and improve user satisfaction. Hence, it is necessary to construct an automatic sentiment analyser for automatic identification of sentiment polarity of the online product reviews. Traditional lexicon-based approaches used for sentiment analysis suffered from several accuracy issues while machine learning techniques require labelled training data. This paper introduces a hybrid sentiment analysis framework to bond the gap between both machine learning and lexicon-based approaches. A novel tunicate swarm algorithm (TSA) based feature reduction is integrated with the proposed hybrid method to solve the scalability issue that arises due to a large feature set. It reduces the feature set size to 43% without changing the accuracy (93%). Besides, it improves the scalability, reduces the computation time and enhances the overall performance of the proposed framework. From experimental analysis, it can be observed that TSA outperforms existing feature selection techniques such as particle swarm optimization and genetic algorithm. Moreover, the proposed approach is analysed with performance metrics such as recall, precision, F1-score, feature size and computation time.

Key words: Tunicate Swarm Algorithm, Feature optimization, sentiment analysis, classifier, machine learning.

AMS subject classifications. 68T05

1. Introduction. The enormous growth of social networks including blogs, forum discussions and e-commerce sites are used among people to share their opinion about online products, services or any topics. These online reviews can be used by many organizations to enhance customer satisfaction, to improve product quality and identify the aspects of products to be upgraded. It also helps to make better-informed decisions towards user interest and preference thereby generating more profits. For example, Amazon collects customer reviews about the products or services. Similarly, social networks like Twitter and Facebook allow their users to share their opinions on any topic like events, elections, products or services. These opinions are useful to manufacturers as well as customers [1]. The manufacturers collect the negative reviews from the customers and rectify them to increase the sales. This kind of analysis is usually called as sentiment analysis (SA) termed as the computational opinion study for several events, topics, products, entities and their qualities. Hence, it is necessary to construct a sentiment analyser for the classification of product data into negative or positive sentiments and for automatic identification of product aspect sentiments from the review documents.

Based on the mechanisms used, SA is mainly classified under three forms like machine learning (ML), lexicon-based and hybrid [2]. In ML-based approaches, various learning algorithms and labelled datasets are utilized to train the classifier for identifying the sentiments [3]. In lexicon-based approaches, the sentiment polarity of the dataset is calculated through the semantic orientation of words [4]. It will classify the texts using unlabelled training set where the lexicons are defined independently of the text; so that overfitting at any instance can be prevented [28]. It showed robust performance across texts and domains and can be applied to perform SA on multiple domain datasets [6]. Though this approach posses more merits, the need for manual maintenance and less accuracy becomes a major disadvantage. Moreover, it could not identify abbreviations in a non-standard form which are mainly used in posts due to limited coverage on informal texts [7]. On the other hand, the ML approach is very suitable for informal text and unstructured content. It provides

*School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, India (danny02.20099@gmail.com).

†School of Computing Science and Engineering, Vellore Institute of Technology, Chennai Campus, India (janakimeena.m@vit.ac.in).

more flexibility and hence eliminates predefined lexicons. Although ML approach performs well, it requires a manually annotated training dataset [8]. As a result, hybrid methods are introduced to bridge the gap between ML and lexicon-based approaches [9]. In this paper, valence aware dictionary for sentiment reasoning (VADER) is utilized for lexicon-based approaches and linear support vector machine (LSVM) classifier is utilized for ML approaches.

Since high-quality word embeddings can be obtained from Glove [10] and fastText [11] models, the proposed work utilizes these models to get initial word embeddings. However, including all the possible features will grow the feature size that would not fit in the memory and cause scalability problem. An approach with better scalability will be able to maintain its performance even with larger datasets. This is usually achieved by adopting suitable feature selection methods [12]. The main aim of feature selection is to generate a well-selected subset of feature that can improve the performance of classifier with better scalability and minimize the time cost simultaneously. Conversely, feature selection through optimization algorithms can automatically select features without manual intervention and eliminate large quantity of unnecessary features without compromising the accuracy.

Conventional feature selection techniques based on concepts like data pick up, shared data and Chi-square are better in reducing the redundant features but have less accuracy. Furthermore, the unwanted features create a non-polynomial hard issue which reduces the system efficiency on selection [13]. Therefore, the attention is now been shifted to intelligent algorithms to enhance classification and to solve the scalability issues. These intelligent algorithms are nature-inspired algorithms that are effective for solving complex problems during classification [14]. It reduces any kind of problems related to accuracy, rate of misclassification and error. Particle swarm optimization (PSO) [15] and genetic algorithm (GA) [16] are two traditional algorithms employed to resolve many complex problems. PSO is influenced by two notable drawbacks such as diversity loss and outdated memory [17]. This can be overcome with a novel nature-inspired algorithm called tunicate swarm algorithm (TSA) [18]. It is effective with redundant data and selects the appropriate features within least execution time.

The main aim of this paper is to propose a hybrid SA method by combining lexicon-based and ML approaches. The data is initially labelled by using VADER sentiment lexicon and the labelled data input to ML classifier. Feature selection is carried out with novel tunicate swarm algorithm. Experiments are performed to show the performance of proposed method with Amazon product reviews from four categories such as Electronics, Toys, Furniture and Camera. The performance is evaluated in terms of metrics such as F1 score, recall, precision and accuracy. Moreover, the computation time comparison of proposed TSA based feature selection with existing methods like GA and PSO is also provided. The experimental results of the proposed method with Amazon dataset shows best classification performance in each test data while adapting more training data. The rest of this paper is arranged as follows. Section 2 details the related works in the area of hybrid SA and optimized feature selection. Section 3 explains the proposed methodology. Experiments are provided in Sect. 4. In Sect. 5, the results of experiments are discussed. Section 6 concludes this paper with future work.

2. Related Works. In this section, the related works carried out in the area of lexicon-based approach, ML based approach and hybrid approach with feature selection are discussed.

2.1. Hybrid Sentiment Analysis. The lexicon-based sentiment analysis methods are widely classified into two approaches: (i) dictionary based and (ii) corpus based [19]. In dictionary-based approach, the words with their polarity scores are utilized for sentiment analysis. In corpus-based approach, positive and negative set of words together with their probability of sentiment is utilized. VADER is a lexicon as well as rule-based SA framework that performs equally or better when compared to existing SA lexicons [20]. Anton and Martin [21] applied the VADER sentiment lexicon with support vector machine (SVM) to classify the sentiments of customer response. Their method is effective with mean AUC of 0.896 and F1 score of 0.834. Tanjim et al. [22] proposed a supervised learning method for SA. Five various ML methods are utilized by them to train the classifier for Amazon product reviews and attained classification accuracy up to 94.02%. When comparing different ML classifiers, LSVM shows the best performance closely followed by random forest (RF). A similar comparison is presented in the proposed work for classifier performance evaluation and accurate measurement.

In lexicon-based methods, the polarities and frequencies of the negative and positive words are examined to get the sentiment of analysed text using a predefined dictionary of words [23]. In ML based approach,

features are generated from the text and used by different learning algorithms to predict a label. A hybrid approach is necessary to eliminate the disadvantages and to combine the merits of each methodology. The hybrid classifier is designed by using ML and lexicon-based methods to classify the documents and to improve the sentiment classification. Term polarities from lexicons are utilized as extra features to train ML classifiers in hybrid approaches [23]. Combination of these methods helps to improve the result of classifier. Xia. et al. [24] presented a hybrid approach with the combination of both ML and lexicon-based methods for SA. Kang et al. [25] introduced a combination of naive bayes (NB) and lexicon-based method for SA of reviews of restaurant. Govindrajan [26] implemented classification-based hybrid SA approach with arcing classifier. The performance of both NB and GA ensemble classifiers are analysed in terms of accuracy. Geetika et al. [27] presented set of approaches to ML with lexical analysis for the product reviews and sentence classification.

A hybrid SA technique on Facebook to automatically analyse sentiments of online product reviews has been proposed by Ortigosa et al. [28]. Zou et al. [29] introduced a method for finding imbalanced threshold classification. Agrawal and Nandi [30] proposed a layered hybrid method for SA. It has two layers where the primary layer is based on the lexicon-based method and the secondary layer is based on the ML method. Rajganesha et al. [31] introduced a hybrid method for SA of feedback-based recommendation system. A hybrid method has been implemented for SA of product reviews from Amazon [32]. In [33], a hybrid SA approach has been introduced by integrating both ML and polarity-based lexicon methods. But these existing works did not employ any optimization algorithms in the feature selection task to improve the scalability and accuracy of hybrid SA.

2.2. Optimized Feature Selection in SA. Feature selection in sentiment analysis plays a vital role in system accuracy enhancement. Many studies have been done in text classification domain with optimized feature selection. Bio-inspired algorithms have an incredible ability in solving different optimization problems [34]. Kalarani et al. [35] proposed the firefly algorithm (FA) for feature reduction with SVM and ANN classifier to classify the sentiments of movie reviews. The experimental outcomes displayed an improved accuracy with reduced training time. Kristiyanti et al. [36] implemented three algorithms called principal component analysis (PCA), PSO and GA based feature reduction with SVM classifier. The accuracy of SVM classifier has been enhanced with these three algorithms. They used Amazon products' reviews for classification. When compared with GA and PCA, PSO algorithm showed higher accuracy for classification.

Farkhund et al. [1] introduced a novel GA based feature selection technique to enhance the scalability issue that happens when the feature-set size increases in the SA. A hybrid SA approach with reviews from IMDB, Yelp and Amazon is utilized. The performance is analysed in terms of recall, accuracy, precision, F1 score and six different classifier algorithms. The results proved that GA based feature selection approach obtains higher accuracy than LSA (Latent semantic analysis) and PCA methods. In the proposed work, a hybrid framework is utilized for SA with novel TSA based optimal feature selection for testing ML classifiers. In sentiment classification, PSO with SVM classifier is the most commonly used approach among researchers followed by ant colony optimization (ACO) [34]. But PSO based optimization has some disadvantages such as the need for multi-objective optimization, fine-tuning of parameters value and poor performance for datasets of multi-domain [37].

This paper proposes a novel TSA based feature reduction in SA. The proposed feature selection approach is encouraged by swarm and jet propulsion behaviours of tunicates throughout the foraging and navigation process [18]. TSA can reduce local optima problem and deliver fast convergence speed by providing better performance in exploration and exploitation stages. Furthermore, it has the capability to keep the stability between exploitation and exploration by searching the large space to get the best global solution. Thus, TSA will be a better solution for feature selection tasks to overcome

3. Proposed Methodology. This section explains the hybrid SA with TSA based optimized feature selection approach. The proposed methodology with VADER sentiment lexicon and ML classifiers is detailed. Then TSA based optimized feature selection is explained

3.1. Hybrid SA with Optimized Feature Selection Approach. This hybrid method includes lexicon-based dictionary for training the ML classifier and bag-of-words as features for testing the ML classifier with TSA based feature selection. A complete framework of proposed work is displayed in Fig. 3.1.

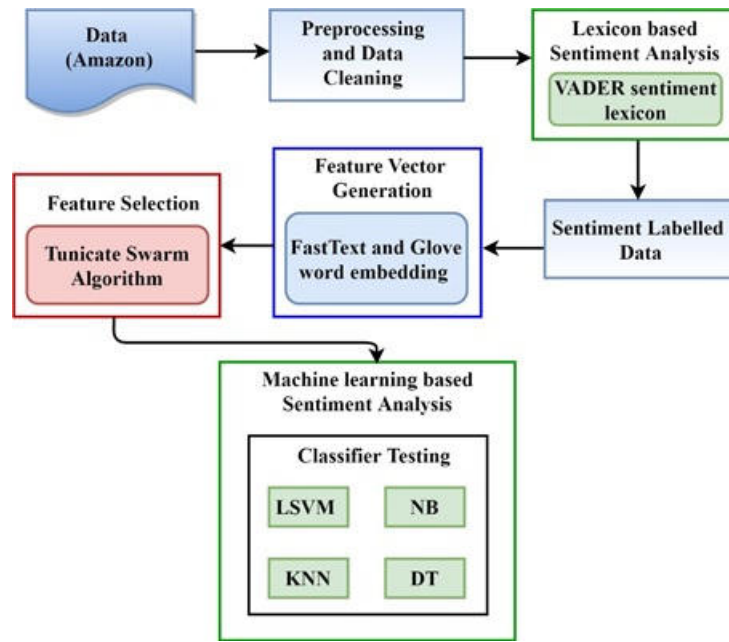


Fig. 3.1: Block diagram of proposed hybrid SA methodology

In the proposed method, the Amazon product data is utilized for SA. Initially, the dataset is pre-processed to eliminate unnecessary data. VADER sentiment lexicon is performed to label the pre-processed data. The feature vectors are generated for labelled training data using Glove and fastText word embeddings. If all the feature vectors are directed to the ML classifier, scalability issue may arise because these vectors contain 80% of the input data. As the dataset size grows bigger, this problem worsens. To minimize this problem, an efficient TSA based feature selection process is introduced in this paper. In this method, each feature vector is modelled on TSA for several hundred generations. TSA simulation is run to find the optimal feature vectors to give improved accuracy for SA. These selected features are utilized to train the ML based classifiers. Linear Support Vector machine, Decision Tree (DT), Naïve Bayes and k-Nearest Neighbours (KNN) are the ML classifiers utilized for classification in this proposed work. The description of these classifiers is given below. Manually labeled dataset is utilized to validate the performance of proposed hybrid SA model.

LSVM. SVM is one of the most used supervised learning techniques which can perform regression, and classification. However, SVM is widely utilized for classification in ML. SVM can handle simple and complex datasets with higher accuracy than other algorithms. In classification, SVM transforms the data points and find hyperplane with maximum margin from multiple decision boundaries to classify the data points in n-dimensional space using the kernel trick technique. The polynomial, linear or Gaussian RBF kernel can be utilized to minimize the computational complexity related to the prediction of new data point. Training an SVM with a linear kernel is quicker than with any other kernel. **NB:** It is also a supervised learning technique mainly utilized in text classification. NB classifiers are based on the Bayes' theorem and occurrence of a particular feature is independent of occurrence of other features. It is a probabilistic based machine learning model that predicts based on the probability of each feature. NB is capable of handle high dimensional complex datasets.

KNN. It is also one of the simplest supervised learning techniques that is utilized for both regression and classification. To classify new data this algorithm assumes the similarity between training data and new data. Then classifies the data based on the similarity. The data is assigned to the category for which the similarity is maximum. To assume the similarity, the distance between training data and testing data is measured by using Euclidean distance. The number of nearest neighbours is calculated for test data. The category of new data is

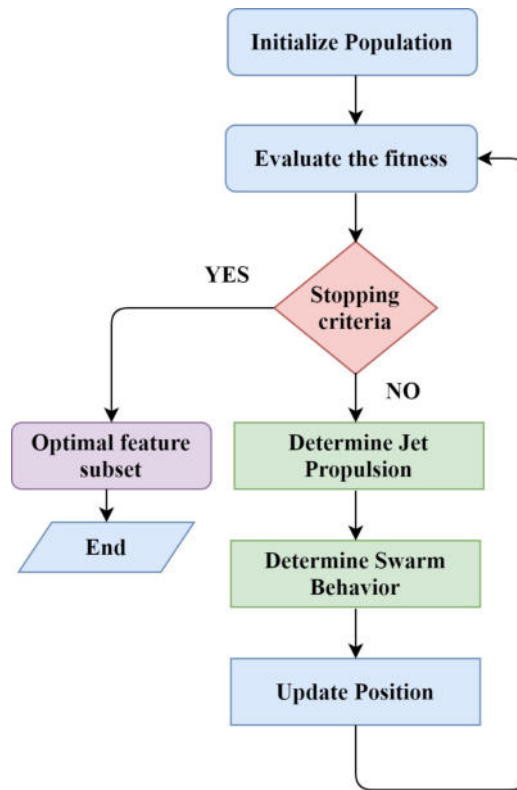


Fig. 3.2: Block diagram of proposed hybrid SA methodology

selected for which data have a greater number of nearest neighbours.

DT. It is also a supervised learning technique that can be utilized to perform both regression and classification. DT utilizes tree structure to classify the data based on given conditions in which root node represents the whole training dataset, decision rules such as Boolean function are represented as branches and output class label is represented by each leaf node. The DT algorithm starts with root node which contains whole dataset. The best attributes are selected using the Attribute Selection Measure. The decision node is created with best attributes. This process is repeated until finding the leaf node for all branches.

3.2. Feature Selection using TSA. When using a larger dataset, scalability issue arises as the increase in the feature vector size. This issue can be eliminated with feature vector optimization by minimizing its size when keeping the accuracy without reduction. This problem is framed in this section and its solution is proposed by utilizing TSA approach.

3.2.1. Problem formulation. If we include all the possible features in the testing set, then the increase in the size of datasets will create larger feature size that creates scalability problem. Thus, to eliminate this problem, the feature selection optimization technique is required. An optimization problem is the minimum number of selection of features from the feature-set of large size. As discussed before, the TSA has more merits such as achieving optimized fitness value and early convergence compared to other optimization algorithms like PSO and GA. Also, this optimization algorithm does not need labelled dataset for sentiment classification.

3.2.2. Mathematical model and optimization. From the training data, the feature vectors are extracted and the optimal feature subset is selected by using TSA approach. Fig. 3.2 shows the flow chart of the proposed TSA based feature selection approach.

In the proposed approach, the optimal feature subset is selected according to the position of best features

(\overrightarrow{FS}) . The features are initialized as the tunicate population $(\overrightarrow{P_p})$. The features can keep its position towards the best feature. The best feature is explored after the computation of fitness value of each feature. In the proposed approach, the feature subset having minimum error rate in prediction of sentiment polarity is considered as the best feature. The fitness value is calculated using Eq. 3.1.

$$(3.1) \quad fitness = \min(1 - accuracy)$$

Here, accuracy in the prediction of sentiment polarity is considered to calculate the error rate. Algorithm 3.1 explains the fitness function calculation.

Algorithm 1: Fitness calculation

```

Procedure Compute Fitness  $(\overrightarrow{P_p})$ 
  for  $i \leftarrow 1$  to  $N$  do
     $\lfloor$  Fit  $[i] = \text{FitnessFunction}(P_p)(\overrightarrow{i}, :)$ 
     $\lfloor$   $Fit_{best} \leftarrow \text{Best}(Fit[])$ ;
  return  $Fit_{best}$ 

Procedure Procedure Best (Fit)
   $Best \leftarrow Fit[0]$ ;
  for  $i \leftarrow 1$  to  $N$  do
    if  $(Fit[i] < Best)$  then
       $\lfloor$   $Best \leftarrow Fit[i]$ 
  return  $Best$ 
  
```

The position of best feature (\overrightarrow{FS}) is explored after the computation of fitness value. (\overrightarrow{PD}) is defined as the distance between the features and the position of best feature. It can be determined using Eq. 3.2.

$$(3.2) \quad (\overrightarrow{PD}) = |(\overrightarrow{FS}) - r_{and} \cdot (\overrightarrow{P_p}(x))|$$

where x represents the present iteration, (r_{and}) represents a random number in the range between 0 and 1. (\overrightarrow{A}) is calculated using Eqs. 3.3, 3.4 and 3.5.

$$(3.3) \quad \overrightarrow{A} = \frac{\overrightarrow{G}}{\overrightarrow{M}}$$

$$(3.4) \quad \overrightarrow{G} = c_2 + c_3 - \overrightarrow{F}$$

$$(3.5) \quad \overrightarrow{F} = 2 \cdot c_1$$

Here \overrightarrow{G} represents the force of gravity and \overrightarrow{F} represents water flow advection in deep ocean [17]. c_1, c_2 and c_3 are random number variables between the range $[0, 1]$. \overrightarrow{M} shows the tunicates social forces. \overrightarrow{M} is calculated as shown in Eq. 3.6.

$$(3.6) \quad \overrightarrow{M} = \lfloor P_{min} + c_1 \cdot P_{max} - P_{min} \rfloor$$

where P_{min} and P_{max} shows the initial and secondary speeds to create social interaction. P_{min} and P_{max} values are considered as 1 and 4 respectively. $(\overrightarrow{P_p}(x))$ is the position of feature which is calculated using Eq. 3.7.

$$(3.7) \quad \overrightarrow{P_p}(x) = \begin{cases} \overrightarrow{FS} + \overrightarrow{A} \cdot \overrightarrow{PD}, & \text{if } r_{and} \geq 0.5 \\ \overrightarrow{FS} - \overrightarrow{A} \cdot \overrightarrow{PD}, & \text{if } r_{and} < 0.5 \end{cases}$$

The position of each features $\overrightarrow{P_p(x)}$ is updated according to the position of best feature $\overrightarrow{F\hat{S}}$ using Eq. 3.8.

$$(3.8) \quad \overrightarrow{P_p(x+1)} = \frac{\overrightarrow{P_p(x)} + \overrightarrow{P_p(x+1)}}{2 + c_1}$$

The updated features going beyond the optimal feature size are adjusted. The fitness value is calculated for the updated position of features. ($\overrightarrow{P_p}$) is updated till a solution is found better than the prior optimal solution. This process is repeated until the satisfied stopping criterion is obtained. Thus, the best optimal feature subset is obtained from the OBL-TSA based feature selection process. Algorithm 3.2 explains the optimal feature subset selection with OBL-TSA based feature selection process.

Algorithm 2: Feature selection with Tunicate Swarm Algorithm

input : Feature set as tunicate population $\overrightarrow{P_p}$, N dimension of features.

output: Position of optimal best features $\overrightarrow{F\hat{S}}$

Procedure *Procedure TSA_feature_selection*

```

//Initialization;
Initialize population  $\overrightarrow{P_p}$ ;
Initialize the parameters  $\overrightarrow{A}$ ,  $\overrightarrow{G}$ ,  $\overrightarrow{F}$ ,  $\overrightarrow{M}$ , and  $Max_{iter}$ ;
Set  $\overrightarrow{P_p} \leftarrow 0$ ,  $P_{min} \leftarrow 1$ ,  $P_{max} \leftarrow 4$ ;
while  $x < Max_{iter}$  do
    //Compute fitness;
    Calculate position of best features  $\overrightarrow{F\hat{S}}$  using Eq. 3.1;
     $\overrightarrow{F\hat{S}} \leftarrow ComputeFitness(\overrightarrow{P_p})$  using;
    // Jet propulsion and swarm behaviour;
     $c_1, c_2, c_3, r_{and} \leftarrow Rand()$ ;
    Determine  $\overrightarrow{P\hat{D}}$  using Eq. 3.2,3.3,3.4,3.5 and 3.6.;
    Calculate the position of features  $\overrightarrow{P_p(x)}$  using Eq. 3.7;
    Update the position of features using Eq. 3.8;
return  $\overrightarrow{F\hat{S}}$ 

```

4. Experiments. This section details which dataset used in this work, how dataset is prepared for SA by removing the redundant data and also explains how to evaluate the proposed model for SA tasks.

4.1. Dataset. The dataset utilized for this experiment is Amazon product reviews from four categories such as Electronics, Toys, Furniture and Camera. It consists of approximately 48,500 product reviews extracted from Amazon.com (<https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt>). The dataset after manual data cleaning process has been taken for analysis. This contains 1000 positive reviews and 1000 negative reviews for each category. Each review has information with rating (0–5 stars), reviewer location and name, review date and title, product name and the review text. The dataset was unlabelled and labelled manually while using as testing data. For that purpose, the ratings of product reviews have been divided into two categories, reviews with ratings >3 labelled as positive and the reviews with rating <3 labelled as negative. These reviews go through data cleaning and pre-processing before given as input to ML classifier because of its unstructured format.

4.2. Data Pre-Processing and Cleaning. In this section, product reviews after manual processing are kept in the memory for pre-processing and cleaning. Pre-processing of data removes unwanted data including web addresses, URLs and online links. This module also includes tokenization and case conversion.

- Removal of unwanted data. This step removes unwanted non-characters consisting of URLs, symbols, web addresses, digits and online links from the review.

Table 5.1: Model Parameters

Algorithm	Parameters	Values
PSO	Weight	0.2
	Constant	2
GA	Rate of Crossover	0.8
	Rate of Mutation	0.01
Proposed TSA	P_{min}	1
	P_{max}	4

- Removal of stop words. Most of the more regularly used stop words in English are “an”, “of”, “a”, “you”, “the”, “and”, etc. These are some words that do not have any meaning. So, these words are generally ignored to improve the accuracy of SA. These words are collected together and removed from the dataset.
- Tokenization. In this process, the sequence of strings is separated into individuals such as keywords, words, symbols, phrases and tokens based on the space of separation. Further, punctuation marks are discarded in this process.
- Case conversion. All the reviews should be converted to lower case because it must be in same case to process. At last, a string of meaningful words is obtained.

4.3. Performance Metrics. In this paper, four performance measures were utilized for the performance evaluation of proposed approach: precision, accuracy, recall and F1 score. The performance measures are given as follows.

- True Positives. It is the count of positive comments classified correctly.
- False Positives. It is the count of negative comments classified wrongly.
- True Negatives. It is the count of negative comments classified correctly.
- False Negatives. It is the count of positive comments classified wrongly.
- Accuracy. It is the ratio between the correctly classified comments and the total number of comments as shown in Eq. 4.1.

$$(4.1) \quad accuracy = TP + TN / TP + TN + FP + FN$$

- Precision. It is the ratio of properly classified positive comments over the total number of positive classified comments as shown in Eq. 4.2.

$$(4.2) \quad precision = TP / TP + FP$$

- Recall. It is the ratio of properly classified positive comments over all comments actually belonging to that class as shown in Eq. 4.3.

$$(4.3) \quad recall = TP / TP + FN$$

- F1-Score. It is the weighted average of recall and precision as shown in Eq. 4.4.

$$(4.4) \quad F1 = 2 * precision * recall / precision + recall$$

5. Parameter Settings. The experiments are performed on a HP laptop with Windows 10 operating system, Intel Core i3 processor having 2.3 GHz frequency, 4GB of RAM. The software used for evaluation of the proposed framework is MATLAB R2020a. The model parameters used to validate the performance of various optimization techniques in SA oriented feature selection domain and general feature selection domain is given in Table 5.1.

Table 6.1: Performance measures comparison of different classifiers in sentiment analysis

Classifier	Accuracy	Precision	Recall	F1 score	Computation Time (sec)
LSVM	93	94.3	97.6	96	22
NB	85	89.7	94.3	92	28
DT	82	88.8	88.4	88.6	32
KNN	87	92.9	96.5	94.7	25

Table 6.2: Performance measures comparison of VADER lexicon model

Products	Accuracy	Precision	Recall	F1 score
Furniture	68	71.3	65.3	68.3
Toys	62	63.5	56.8	59.9
Electronics	63	65.3	57.7	61.5
Camera	59	61	58.4	59.7

6. Experiment Results and Discussion. The experiment results and discussion are given in this section. The performance of the proposed SA approach is evaluated using four different ML based classifiers: LSVM, NB, DT and KNN. The accuracy, recall, precision and F1 score are used as performance measures for classifiers.

6.1. Performance Comparison of ML classifiers. The comparison of four various ML classifiers in the proposed hybrid SA approach is performed by applying TSA and non-TSA enhanced features. The same tests are performed for the reviews of four different products that have been discussed previously.

Table 6.1 shows the accuracy, precision, recall and F1 score comparison for four different classifiers. These results are taken for SA of Furniture dataset. The accuracy of LSVM is higher than other classifiers. It has achieved more than 90% accuracy which lies below 90% for remaining classifiers. The precision rate of all the classifiers are observed to be more than 85%. However, precision rate of LSVM is comparatively higher than other classifiers. The recall measure and F1-score of all classifiers except DT attained more than 90%. The recall rate is 1% higher for LSVM compared to KNN approach. Likewise, F1-score is slightly better for LSVM in comparison to both NB and KNN. When considering computation time, LSVM takes less time than other classifiers. From these results, it can be observed that total performance of LSVM classifier is better than other classifiers. In the proposed work, LSVM classifier is used for classification of other datasets since it proves best accuracy than other classifiers.

6.2. Comparison of three different SA approaches. Three SA approaches are implemented in the proposed SA framework. Table 6.2 shows the performance comparison for VADER sentiment lexicon model. The accuracy of all datasets except Camera dataset attained more than 60% . The accuracy of Furniture dataset achieved 68% which lies below 65% for other datasets. In terms of precision, the Furniture dataset has the maximum precision of 71.3%. Similarly, the F1 score of Furniture dataset is comparatively higher than other datasets. The Recall rate is 65.3% for Furniture dataset.

Table 6.3 shows the performance measures comparison of ML approach for LSVM classifier with TSA based feature selection approach. The fastText and Glove word embedding model is utilized to create feature vectors. From Table 4, it can be observed that Camera dataset has the maximum accuracy than other datasets. The precision rate of Furniture dataset is comparatively higher than other datasets. In terms of Recall, Camera dataset shows a higher recall than Electronics dataset. The F1 score of all datasets was observed to be more than 70%. The F1score is slightly better for Camera dataset in comparison to Electronics dataset.

Table 6.4 shows the results of the proposed model with TSA and without TSA. LSVM classifier is utilized to classify the Amazon product data as it shows better results. From Table 5, it can be observed that the accuracy of TSA and non- TSA based approaches for Furniture dataset achieved maximum accuracy than

Table 6.3: Performance measures comparison of ML approach with LSVM classifier

Products	Accuracy	Precision	Recall	F1 score
Furniture	79	81.8	62	71.9
Toys	74	75.2	71.6	73.4
Electronics	78	78	78.8	78.4
Camera	85	70.2	87.2	78.7.7

Table 6.4: Performance measures comparison of proposed hybrid model with TSA and without TSA on Amazon data

Products	With TSA				Without TSA			
	Accu racy	Preci sion	Recall	F1 score	Accu racy	Preci sion	Recall	F1 score
Furniture	93	94.4	97.6	96	91	93.9	97.3	95.6
Toys	86	92.5	95.3	93.9	84	89.8	91.6	90.7
Electronics	89	93	96.2	94.6	88	89.3	95.1	92.2
Camera	85	87.5	93.1	90.3	82	85.4	92.8	89.1

other datasets. The precision rate of all datasets achieved above 90% except Camera dataset for both TSA and non- TSA optimization. The Recall and F1 score is more than 90% for all datasets with TSA optimization. Without TSA optimization, the Recall rate is above 90% for all datasets. Likewise, F1 score is above 90% for all datasets except Camera dataset without TSA optimization. These results show that the performance measures of TSA based optimized feature selection is almost similar to non-TSA based technique with reduced feature size. Thus, TSA based optimal feature selection reduces the scalability problem and computation time than non- TSA based approaches. Thus, Tables 3, 4 and 5 shows that the proposed hybrid SA approach with TSA based feature selection outperforms VADER lexicon and ML-based approaches.

6.3. Feature size comparison of TSA based ML approach. Feature size comparison is performed to show that TSA based feature selection improves the scalability by reducing the feature size while keeping the same accuracy as non-TSA based SA.

Figure 6.1 shows the feature size before and after SA optimization on selection of features. For Electronics dataset, the size of features before employing TSA is 3295. After employing TSA, the feature size has been minimized to 2169 which is nearly 34% less in the size. For the Toys dataset, the size of features before employing TSA is 4045. After employing TSA, the feature size has been minimized to 2985 which is almost 26% less in the size. For Camera dataset, the size of features before employing TSA is 2940. After employing TSA, the feature size has been minimized to 1754 which is almost 40% less in the size. Moreover, for Furniture dataset, the size of features before performing TSA is 3492. After performing TSA, the feature size has been minimized to 1982 which is nearly 43% less in the size. We have already seen that the performance of both methods is similar but optimization using TSA gives minimized size of the feature. This reflects an important effect on the system scalability. When dataset with large features is used for SA, it results in an enormous bottleneck.

6.4. Comparing of TSA with GA and PSO. The results are taken only for LSVM classifier because LSVM gives a better outcome for both TSA and non-TSA methods. To make a comparison and to verify the results of TSA, two famous existing algorithms: GA and PSO are considered.

Figure 6.2 illustrates that TSA finds the optimum solution faster than the other two optimization algorithms and it has an improved convergence speed. Thus, the proposed TSA based feature selection algorithm can keep the balance between the exploration and exploitation and escape from local optima problem.

From Table 6.5, it can be known that TSA based feature selection shows better accuracy of 93% which is

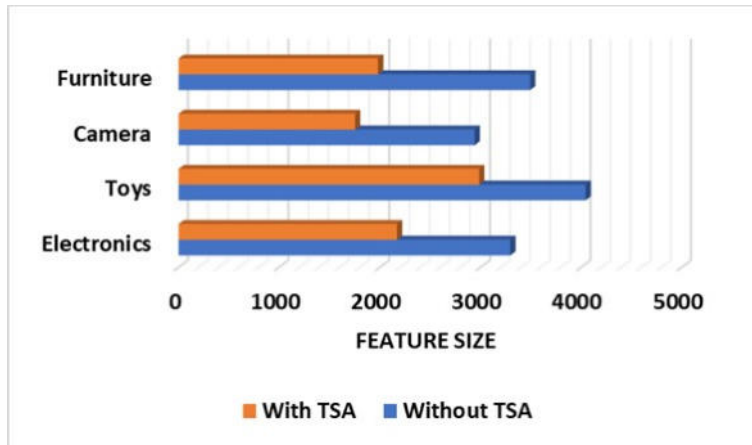


Fig. 6.1: Features size comparison for various datasets with TSA and without TSA

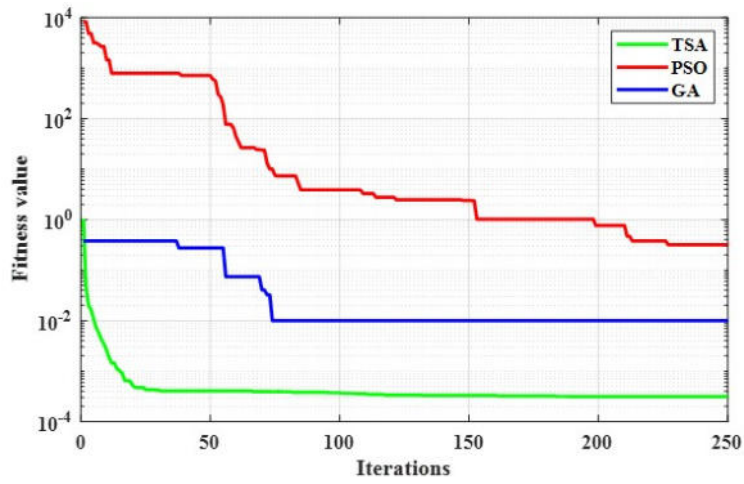


Fig. 6.2: Convergence curve comparison for PSO, GA in general feature selection domain

Table 6.5: Performance Comparison of various feature selection techniques

Techniques	Performance measures					
	Accuracy	Precision	Recall	F1 score	Computation time	Feature size reduction (%)
Proposed OBL-TSA	95	95.4	96.6	96	15 sec	43
TSA	93	94.3	97.6	96	22 sec	40
PSO	89	91.1	96.5	93.7	35 sec	38
GA	90	92.1	96.5	94.3	38 sec	35

Table 6.6: Comparative Analysis with related works

Methods	Dataset	Accuracy (%)
SA with Probabilistic Machine Learning for Amazon Reviews [38]	Books	82.9
	Electronics	86.6
	DVD	83.7
	Kitchen	89.1
SA for business analytics with and cellphone Amazon Reviews [39]	Accessories & Cellphone	80.11 72.95
	SA for Amazon Product Reviews and feature selection and methods [40]	Camera
80		
68		
Musical instruments		62
		80
		68
Books		70
		70
		80
Proposed Model	Electronics	89
	Toys	86
	Camera	85
	Furniture	93

better than PSO and GA based feature selection with less computation time (22 sec). Moreover, 43% of the features are minimized using TSA based SA. The final part of this discussion demonstrates that TSA based feature reduction outperformed both GA and PSO based feature selection approaches. From these results, it can be concluded that the proposed TSA based approach is efficient than other commonly used existing algorithms for feature selection.

Table 6.6 shows the comparative analysis of the proposed method with the related works in terms of accuracy. The related works employed various pre-processing and feature extraction processes. In the proposed approach hybrid approach is employed with optimized feature selection. From this comparative analysis and above results, it can be observed that the proposed approach is more effective than existing optimization algorithms for feature selection and could give better results than related works.

7. Conclusion. In this paper, a novel hybrid model with feature selection using TSA is designed, developed and evaluated. Moreover, the proposed hybrid technique is evaluated for scalability in terms of execution time comparison. In the total execution time, optimal feature-set selection using TSA took about 50-60% and reduced feature-size up to 43% without any change in accuracy (93%). Experiments showed the performance of proposed method with the data of Amazon product reviews from four categories Electronics, Camera, Furniture and Toys. The evaluation of the proposed work is done with performance metrics such as F1-score, recall, precision and accuracy. The results indicated that the TSA based optimized feature selection method showed improved accuracy than PSO and GA algorithms with less computation time. Thus, the proposed approach has higher accuracy and better scalability for SA of online reviews. The future work aims at the extension of the proposed work on multi-domain SA with various sources of datasets.

REFERENCES

- [1] F. IQBAL, J. MAQBOOL, B. C. M. FUNG, R. BATOOL, A. M. KHATTAK, S. ALEEM AND P. C. K. HUNG, *A Hybrid Framework for Sentiment Analysis using Genetic Algorithm based Feature Reduction*, IEEE Access, 1-1, 2019.

- [2] S. KAUSAR, X. HUAHU, M. Y SHABIR AND W. AHMAD, *A Sentiment Polarity Categorization Technique for Online Product Reviews*, IEEE Access, 8, 3594–3605, 2020.
- [3] E. BOIY AND M.-F. MOENS, *A machine learning approach to sentiment analysis in multilingual Web texts*, Information Retrieval, 12(5), 526–558, 2008.
- [4] M. TABOADA, J. BROOKE, M. TOFILOSKI, K. VOLL AND M. STEDE, *Sentiment analysis in Facebook and its application to e-learning*, Computers in Human Behavior, 37(2), 267–307, 2014.
- [5] A. ORTIGOSA, J.M. MARTÍN AND R.M. CARRO, *Lexicon-Based Methods for Sentiment Analysis*, Computational Linguistics, 31, 527–541, 2011.
- [6] G. KATZ, N. OFEK AND B. SHAPIRA, *ConSent: Context-based sentiment analysis*, Knowledge-Based Systems, 84, 162–178, 2015.
- [7] P. DUCANGE, M. FAZZOLARI, M. PETROCCHI AND M. VECCHIO, *An effective Decision Support System for social media listening based on cross-source sentiment analysis models*, Engineering Applications of Artificial Intelligence, 78, 71–85, 2019.
- [8] Z. YUAN, S. WU, F. WU, J. LIU AND Y. HUANG, *Domain attention model for multi-domain sentiment classification*, Knowledge-Based Systems, 155, 1–10, 2018.
- [9] L. SHANG, Z. ZHOU AND X. LIU, *Particle swarm optimization-based feature selection in sentiment classification*, Soft Computing, 20(10), 3821–3834, 2016.
- [10] J. PENNINGTON, R. SOCHER AND C. MANNING, *Glove: Global Vectors for Word Representation*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543, 2014.
- [11] P. BOJANOWSKI, E. GRAVE, A. JOULIN AND T. MIKOLOV, *En-riching Word Vectors with Subword Information*, Transactions of the Association for Computational Linguistics, 5, 135–146, 2017.
- [12] G. ANSARI, T. AHMAD AND M.N. DOJA, *Hybrid Filter-Wrapper Feature Selection Method for Sentiment Classification*, Arabian Journal for Science and Engineering, 2019.
- [13] Y. LIU, C. GOA, Z. ZHANG, Y. LU, S. CHEN, M. LIANG AND L. TAO, *Solving NP-Hard Problems with Physarum-Based Ant Colony System*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(1), 108–120, 2017.
- [14] A. NABAEI, M. HAMIAN, M.R. PARSAEI, R. SAFDARI, T. SAMAD-SOLTANI, H. SAMAD-SOLTANI AND A. GHASSEMI, *Topologies and performance of intelligent algorithms: a comprehensive review*, Topologies and performance of intelligent algorithms: a comprehensive review, Artificial Intelligence Review, 49(1), 79–103, 2016.
- [15] R. EBERHART AND J. KENNEDY, *A new optimizer using particle swarm theory*, MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science.
- [16] J.H. HOLLAND, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, U Michigan Press.
- [17] S.S. GRILL, R. BUYYA, I. CHANA, M. SINGH AND A. ABRAHAM, *BULLET: Particle Swarm Optimization Based Scheduling Technique for Provisioned Cloud Resources*, Journal of Network and Systems Management, 26(2), 361–400, 2017.
- [18] S. KAUR, L.K AWASTHI, A.L. SANGAL AND G. DHIMAN, *Tunicate Swarm Algorithm: A new bio-inspired based metaheuristic paradigm for global optimization*, Engineering Applications of Artificial Intelligence, 90, 103541, 2020.
- [19] M. DARWICH, S.A.M. NOAH, N. OMAR AND N. OSMAN, *Corpus-Based Techniques for Sentiment Lexicon Generation: A Review*, Journal of Digital Information Management, 17, 296, 2019.
- [20] C.J. HUTTO AND E. GILBERT, *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM, 2014.
- [21] A. BORG AND M. BOLDT, *Using VADER Sentiment and SVM for Predicting Customer Response Sentiment*, Expert Systems with Applications, 113746, 2020.
- [22] T.U. HAQUE, N.N. SABER AND F.M. SHAH, *Sentiment analysis on large scale Amazon product reviews*, IEEE International Conference on Innovative Research and Development (ICIRD), 2018.
- [23] H. CHO, S. KIM, J. LEE AND J.S. LEE, *Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews*, Knowledge-Based Systems, 71, 61–71, 2014.
- [24] R. XIA, C. ZONG AND S. LI, *Ensemble of feature sets and classification algorithms for sentiment classification*, Information Sciences, 181(6), 1138–1152, 2011.
- [25] H. KANG, S.J. YOO AND D. HAN, *Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews*, Expert Systems with Applications, 39(5), 6000–6010, 2012.
- [26] M. GOVINDARAJAN, *Sentiment analysis of movie reviews using hybrid method of naive Bayes and genetic algorithm*, Int. J. Adv. Comput. Res., 3(4), 139145, 2013.
- [27] G. GAUTAM AND D. YADAV, *Sentiment analysis of twitter data using machine learning approaches and semantic analysis*, Seventh International Conference on Contemporary Computing (IC3), 2014.
- [28] A. ORTIGOSA, J.M. MARTÍN AND R.M. CARRO, *Sentiment analysis in Facebook and its application to e-learning*, Computers in Human Behavior, 31, 527–541, 2014.
- [29] Q. ZOU, S. XIE, M. WU AND Y. JU, *Finding the Best Classification Threshold in Imbalanced Classification*, Big Data Research, 5, 2–8, 2016.
- [30] V. NANDI AND S. AGRAWAL, *Political sentiment analysis using hybrid approach*, International Research Journal of Engineering and Technology (IRJET), 3(5), 1621–1627, 2016.
- [31] N. RAJGANESH, C. ASHA, A.T. KEERTHANA AND K. SURIYA, *A hybrid feedback-based book recommendation system using sentiment analysis*, IJSRCSEIT, 3(3), 2456–3307, 2018.
- [32] D. MUMTAZ AND B. AHUJA, *A Lexical and Machine Learning-Based Hybrid System for Sentiment Analysis*, Studies in Computational Intelligence, 165–175, 2017.
- [33] I. GUPTA AND N. JOSHI, *Enhanced Twitter Sentiment Analysis Using Hybrid Approach and by Accounting Local Contextual Semantic*, Journal of Intelligent Systems, 0(0), 2019.

- [34] A. YADAV AND D.K. VISHWAKARMA, *A comparative study on bio-inspired algorithms for sentiment analysis*, Cluster Computing, 2020.
- [35] P. KALARANI AND S. BRUNDA, *Sentiment analysis by POS and joint sentiment topic features using SVM and ANN*, Soft Computing, 2018.
- [36] D.A. KRISTİYANTI AND M. WAHYUDI, *Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review*, 5th International Conference on Cyber and IT Service Management (CITSM), 2017.
- [37] H. JIANG, C.K. KWONG, W.Y. PARK AND K.M. YU, *A multi-objective PSO approach of mining association rules for affective design based on online customer reviews*, Journal of Engineering Design, 29(7), 381–403, 2018.
- [38] R. XIA, F. XU, J. YU, Y. QI AND E. CAMBRIA, *Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis*, Information Processing & Management, 52(1), 36–45, 2018.
- [39] M.S. ELLI AND YI-FAN, *Amazon Reviews, business analytics with sentiment analysis*, 2016.
- [40] T. SHAIKH AND D. DEEPA, *Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews*, International Journal of Computer Trends and Technology, 36(4), 225-230, 2016.

Edited by: Dana Petcu

Received: Nov 26, 2020

Accepted: Jan 23, 2021