IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.Doi Number

GAWA – A Feature Selection Method for Hybrid Sentiment Classification

A. Rasool[1,2], R. Tao[1], M. Kamyab[1], H. Shoaib[1].

[1]College of Computer Science and Technology, Donghua University, Shanghai, 201620, China.
[2]Shenzhen Key Lab for High Performance Data Mining, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.

Corresponding author: A. Rasool (e-mail: 317089@mail.dhu.edu.cn).

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant No. 19D111201 and by the Key-Area Research and Development Program of Guangdong Province under Grant No. 2019B010137002.

ABSTRACT Sentiment analysis or opinion mining is the key to natural language processing for the extraction of useful information from the text documents of numerous sources. Several different techniques, i.e., simple rule-based to lexicon-based and more sophisticated machine learning algorithms, have been widely used with different classifiers to get the factual analysis of sentiment. However, lexicon-based sentiment classification is still suffering from low accuracies, mainly due to the deficiency of domain-oriented competitive dictionaries. Similarly, machine learning-based sentiment is also tackling the accuracy constraints because of feature ambiguity from social data. One of the best ways to deal with the accuracy issue is to select the best feature-set and reduce the volume of the feature. This paper proposes a method (namely, GAWA) for feature selection by utilizing the Wrapper Approaches (WA) to select the premier features and the Genetic Algorithm (GA) to reduce the size of the premier features. The novelty of this work is the modified fitness function of heuristic GA to compute the optimal features by reducing the redundancy for better accuracy. This work aims to present a comprehensive model of hybrid sentiment by using the proposed method, GAWA. It will be valued in developing a new approach for the selection of feature-set with a better accuracy level. The experiments revealed that these techniques could reduce the feature-set up-to 61.95% without negotiating the accuracy level. The new optimal feature sets enhanced the efficiency of the Naïve Bayes algorithm up to 92%. This work is compared with the conventional method of feature selection and concluded the 11% better accuracy than PCA and 8% better than PSO. Furthermore, the results are compared with the literature work and found that the proposed method outperformed the previous research.

INDEX TERMS Feature selection, Genetic algorithm, hybrid sentiment classification, machine learning algorithms, Wrapper approach.

I. INTRODUCTION

Online Twitter users share their emotions such as joy and sorrow about any product or activity. Other users can have better knowledge through the existed reviews by the users who have experience with specific items [1]. Forbes reported that 2.5 quintillion bytes of data are being generated every day [2]. In business analytics, this massive data is worthy, but it contains enormous slangs and redundancy [3]. Similarly, due to the text limitation of the Twitter message (Tweet), there are numerous classification problems for Twitter datasets like grammatical and spelling mistakes, insertion of Emoji, hashtag, and use of multiple languages that must be determined for the better acknowledgment of text analysis [4]. This research is emphasizing the essential need for evaluation of user-generated data to address the issue of detecting, extract, and analyze the user opinions for the progression of organizations more efficiently and effectively [5].

Research communities and various academicians are continuously working on to compete for these text classification issues by an emerging technique of opinion and emotion detection named as sentiment analysis [6]. Sentiment Analysis (SA) is a new research interest for a plethora of real-world applications [7]. It is capable of discovering the opinion information from the online user's data to assist the

stakeholders in making a better future's decisions. Sentiment or opinion classification has an immense impact on multiple fields of life. For example, SA has been used for precision marketing, product quality information, stock market forecast, business decision making, election prediction, and counterterrorism [8].

Generally, there are three methods for SA; machine learning, including supervised practice [9], [10], lexicon-based, including unsupervised methods [11], [12], and hybrid approach containing both supervised and unsupervised method [12], [13]. Most of the researchers [13]–[15] worked over the Machine Learning (ML) approach for the sentiment and text mining, where a labeled dataset is used to train their model [9], [16]. The most typical algorithms for ML are Naïve Bayes (NB) [2] and Support Vector Machine (SVM) [17]. The major issue with ML for sentiment is the development of appropriate datasets to train the specific classifiers according to the domain [18]. In the lexicon approach, one or more than one sentimental dictionary is developed or applied to calculate the sentiment polarity of a specific document or segment. However, in many cases, trusting on word effectiveness is not sufficient for consistent results of sentiment detections [19]. The third emerging technique of sentiment is hybrid, which is a combination of ML and lexicon approaches [20]. The hybrid approach is testified with different methods, for example, Principal Component Analysis (PCA) [21] and Particle Swarm Optimization (PSO) [22] for feature selection to acquire better accuracy. Meanwhile, the results are still appetizing for more accuracy due to the feature ambiguity from various sources.

Feature selection is an adoptive process to tackle the feature ambiguity by finding the ultimate and relevantly essential features [23]. The performance of the machine learning algorithm is severely affected by the enormous features. It is quite challenging to detect the optimum feature sets and omit the noisy one. In social media, heterogeneous datasets, especially Twitter data, has complex relationships or interactions among the features. Similarly, it has several irrelevant and redundant feature sets. Hence, these features are not advantageous and naturally lead to ambiguous classification accuracy. The optimal features set for a learning-based model ought to be a subset of essential features that should be distinguished correctly. Thus, the feature selection technique pursues to enrich the classification accuracy and reduce dimensionality and computational complexity [24].

As far as the accuracy is concerned, it relies on the optimal feature set, which can be achieved by using the Wrapper Approaches (WA) [25]. These features or variables become more worthy when some classifiers are implemented on these feature sets. For this research, a Genetic Algorithm (GA) is executed with a modified fitness function. This algorithm has been adapted in various studies [9], [21], [23], [47], [51], [52] with different methods for the better selection of features. The exciting research on sentiment classification proves that previous researchers are trying to test the different algorithms to customize the preference and parameters to achieve better-optimized results than before. It is evident that still up-to-date, there is a strong need to fill the gap about the accuracy of SA.

This research proposes a method for the implementation of hybrid sentiment classification with the GA and WA. The contribution of this paper is a novel method, by induction of the WA for premier features before the implementation of a GA with modified fitness function for optimal features. This evolutionary algorithm results in the improvement of accuracy besides the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon dictionary. However, a comprehensive comparison is conducted with PCA and PSO to determine the efficiency of this proposed work. Furthermore, this article is supported by a distinguished framework for better acknowledgment of experimental implementation process. The dataset is crawled from the famous microblogging site, Twitter, and pre-processed. Concisely, this article is efficient in answering the question: Which method is the best choice for a data analyst to get the optimal feature sets? The experimental results signify the efficiency of the proposed work. The significant contribution of this research article as follows:

- The novelty of the proposed research is a method (GAWA) for optimal feature selection by modified fitness function of the Genetic algorithm assembled with the Wrapper approaches for premier features.
- In the present works of feature selection with the wrapper approach and genetic algorithm have not discussed how the accuracy of ML algorithm change with the different number of optimal selected features. The GAWA based optimal features will provide this answer by employing its proposed method.
- The GAWA feature's performance is analyzed with ML classifiers and confusion matrices. The accuracy of GAWA is compared with previous related work, and with two major feature-reduction algorithms, PCA and PSO.
- A supported framework is designed for hybrid sentiment classification to reduce the implementation complexity.

The structure of the rest article is as follows; Section II deliberates the related research work of SA, feature selection, GA, WA, etc., Section III introduces its proposed method, Section IV deals with experimental implementation and examines the results, Section V elaborates the conclusion and future work.

II. RELATED WORK

Sentiment analysis or opinion mining is a vital category of text analysis that is capable of extracting, and analyzing the opinioned text by classifying the positive, negative, and neutral entity [9]. This entity can comprise the people, service, product, etc. Sentiment analysis has been adopted in numerous

applications for different data sets, such as Weibe *et al.* [26] has practiced the data about automobiles, travel destinations, movies, and banks reviews and design the classification of words into two classes, i.e., positive and negative. This classification was helpful to improve the effectiveness, but it can't detect the like or dislike of opinion holders about each feature. Xia *et al.* [10] constructed a framework to apply the sentiment classification tasks to integrate the feature sets and classifications algorithms to generate better sentiment classification procedures. The author handled the two different kinds of features sets, namely word-relation and POS-based features for opinion mining by utilizing three simple text classification algorithms, including NB, Maximum Entropy (MaxEnt), and SVM. It concluded the effectiveness of the ensemble technique about sentiment classification, but it could not find which combination of feature sets and algorithms is super useful. In opposition to [10], this proposed research promising the novel features along with a GA classifier as a more effective ensemble technique.

A dozen of research e.g., [1], [2], [5], [11], [12], [16], and [55] have been employed by considering the Twitter data as a source dataset for sentiment and emotion extractions. Still, there is a common problem of accurate sentiment classification. This proposed study is dealing with this issue by employ the annotated text with sentimental based dictionaries. These dictionaries, e.g., the lexicon dictionary, contain a list of the words, and each word is allocated a value which indicates the positivity or negativity level [18]. Numerous dictionaries have been developed automatically or semi-automatically [27], for example, SentiWordNet [28] and WordNet-Affect [29]. But Rakibul *et al.* [27] concluded that the VADER dictionary outperforms the SentiStrength, AFINN, and MPQA dictionary concerning evaluation metrics (Precision, Recall, F-score). However, in the case of feature selection, trusting on word efficiency and performance is not satisfactory for equitable analysis of hybrid sentiment detections [19].

Feature selection is deliberated as a vital part of many machine learning methods that remarkably affect the model accuracy. Generally, it has two contractional aims, maximizing the classification's performance and minimizing the feature size [30], [31]. Several feature selection techniques have been introduced, like Mutual Information (MI), Term Frequency - Inverse Document Frequency (TF-IDF), Information Gain (IG), and Chi-Square (CS) for the sentiment classification with a different machine learning algorithm. [32]. For instance, TF-IDF has been used in [33] for the creation of feature vocabulary with Logistic Regression (LR), NB, Decision Tree (DT), and SVM. Features selections methods are essential to reduce the training time for better performance by eliminating unnecessary features [34]. Sharma et al. [35] proposed a hybrid ensemble learning method for the optimal features sets by utilizing various feature selection algorithm. Liu S. and Fan et al. [36] proposed a selection method for the entity main features, which relies on the quantitative dynamic sensor data. It utilized the feature matrix to remove the inappropriate entity features and employed iRelief algorithm to compute the relevant features. It concluded that average search accuracy was improved by more than 10% by the proposed method. Z. Mingxi and W. Jinhua *et al.* [37] proposed a similarity measure method named as HeteRank for general manner relationships between objects by computing similarities in real heterogeneous datasets. It applied a pruning algorithm to improve the computation scalability by ignoring the redundant actions. The experiments yielded the efficiency and effectiveness of HeteRank. Author [38] proposed an algorithm, SA-Cluster, which is based on structural and attribute similarities to measure the vertex closeness on heterogeneous attributes in large graphs. It used the K-Medoids cluster approach to divide the large graph into k clusters. It provided theoretical analysis to prove SA-cluster is converging, and it compared the cluster quality with S-Cluster and W-Cluster and found the effectiveness of SA-Cluster.

Features selection techniques are primarily considered into filter [39] and Wrapper [40] approaches. The recent literature review presents the Wrapper approach as a better performer for sentiment classification as compared to the filter approach [41]. For instance, Gokalp *et al.* introduced a Wrapper based feature selection algorithm for SA [40]. Similarly, Al-Tashi *et al.* [42] projected a multi-objective method for feature selection and reduction by employing the Wrapper based algorithm to assess the performance of selected features for classification. The Wrapper feature selection approach has been widely used in numerous applications, e.g., in the medical field for the calculation of optimum features from coronary artery disease [43]. The author [44] presented a wrapper approach for sentiment polarity classification by the integration of genetic algorithm and SVM classifier. It concluded that the accuracy of polarity classification had been improved by attaining the optimal features sets from the Internet Movie Database (IMDb). Research conducted by Karegowda *et al.* [45] to solve the computation problem of massive dimensional data to make a reliable classification by Wrapper approach for the precise feature selection. The Wrapper approach works on a greedy search algorithm which evaluates all suitable combination of features and chooses the best set of features for a machine learning algorithm. It consists of three different categories; Forward Feature Construction (FFC), Backward Feature Elimination (BFE), and Exhaustive Feature Selection (EFS).

Recently, researchers have developed a metaheuristic approach to provide a satisfactory solution for feature reduction by ensemble the Genetic Algorithm (GA). Such as, Hammami *et al.* [46] proposed a hybrid algorithm to solve the difficulties of expensive computational behavior of the wrapper approach. It used the Wrapper approach with sorting the Genetic algorithm to reduce the scalability issue and achieved competitive results. However, it suggested that feature size can be further reduced by using the fitness

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3030642, IEEE Access

function of GA. Zhou *et al.* [47] presented work to detect the features from high-dimensional data by using the GA with customized parameters for the selection of maximum convergence values by considering fitness function. It concluded the classification error decreased by increasing the fitness function value. Chakraborty *et al.* [48] utilized the GA's fitness function with fuzzy logic to measure the feature quality that was detected as a subset from the feature selection operator. It resolved the computation time of GA with modified fuzzy-based fitness function into two stages of feature selection and classification.

Govindarajan *et al.* [49] presented a hybrid approach by operating the NB and GA for feature reduction and analyzed the performance of different classifiers to determine the accuracy. It concluded that a hybrid classifier represents a source of accuracy improvement. But his proposed work did not compare the performance of multiple classifiers for optimization of feature reduction. A hybrid genetic algorithm was combined with Particle Swarm Optimization (PSO) to improve the classification by the feature selection method [50]. Zainuddin *et al.* [12] proposed work to reduce the big data dimensions with some feature selection methods such as PCA. It found that PCA eliminates the irrelevant and redundant features to gain a higher accuracy for sentiment classification. At the same time, PSO [22] has been employed in various feature selection issues. In recent researches [58], [59], [60], the LSA and PCA, which is quite similar to LDA, have been used for classification accuracy. In this proposed case, PCA and PSO have been applied to evaluate and compare the results of the proposed technique. One of the comprehensive researches that has been performed by Iqbal *et al.* [21] to solve the accuracy issue for machine learning approaches. It designed a hybrid framework to count better performance and to improve efficiency. It employed the genetic algorithm and succeeded to reduce the 42% of the feature set. Despite this achievement, it faces the accuracy issues that can be improved by utilizing a more precise and optimal feature set.

In a nutshell, all these studies have been engaged efficiently for feature selection with different recipes and techniques. However, these practices delivered some accuracy limitations for the tweeter datasets. In contrast, this proposed work is committed to deploying a novel technique by presenting a method for a better selection of features that will be more capable of enhancing accuracy.

III. PROPOSED METHOD – GAWA

To capture the essential but hidden variables for the significant insight is an endeavor task, but it is the best choice for best decisions and predictions. Most of the data scientists [30], [31] are trying to find the best methods to achieve the best features insights. In this research, the WA is used to select the feature sets (named as premier feature sets) across the given dataset. Furthermore, it implements a GA to extract the optimal features from the previous premier feature sets.

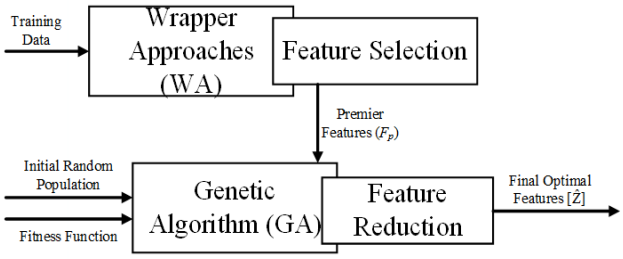

FIGURE 1. Block diagram of the proposed method with the combination of WA and GA.

The relationship between the GA and WA techniques to design this proposed method (GAWA) can be easily understood in Fig. 1.

Algorithm 1 Proposed GAWA Method

Input: Number of features $\sum_{i\leftarrow 0}^{n} f \leftarrow T_f$
Output: Set of optimal features $\mathbb{Z}$
T_f: Total features obtained from the labeled dataset
F_p: Set of all premier features
P: The population of premier features F_P
/* VF_{FFC} is Validated Feature from the FFC technique, and VF_{BFE} is Validated Feature from the BFE technique. */
1: Initiate VF_{FFC} with zero features
2: **for** add features (FFC) **do**
3: **if** accuracy of new added feature > 80% **then**
4: Validate the FFC feature (VF_{FFC})
5: **end if**
6: **end for**
7: **for** remove features (BFE) **do**
8: **if** accuracy of remaining feature > 80% **then**
9: Validate the BFE feature (VF_{BFE})
10: **end if**
11: **end for**
12: $F_p \leftarrow VF_{FFC} + VF_{BFE}$;
13: $P \subseteq F_p$;
14: **if** Fitness Function (FF) of premier features > 0.05
15: Add to convergent population for $\mathbb{Z}$ **else**
16: **while** the termination criterion is not satisfied **do**
17: Perform crossover operator
18: Perform mutation operator
19: Update population for the next generation
20: **end while**
21: **end if**
22: **Return** Optimal Features $\mathbb{Z}$

Algorithm 1 starts by taking the numbers of typical features that have been attained from the labeled dataset. The input is the sum of all features in the form of total feature (T_f) and processes the FFC and BFE approaches to select premier features sets (F_P). A GA-based feature reduction technique is used which utilize the F_p of the Wrapper approach and yield

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3030642, IEEE Access

the set of optimal features $\mathbb{Z}$ that will be trained to assess the machine learning algorithm's performance. In-depth, the working functionality of the proposed algorithm will be elaborated below in subparts of the WA and GA.

A. WRAPPER APPROACHES (WA)

In this research, FFC and BFE have been implemented with the *RandomForestClassifier* algorithm as an estimator. FFC is initiated with zero feature set, and in the next iterations, it continuously adds the feature until it improves the efficiency of the model. It stops automatically when the addition of a new feature minimizing the performance. However, BFE is initiated with all features. In the next iteration, it removes the least essential or redundant features until it starts to provide less accuracy in results. The functionality criteria of FFC and BFE can be clearly understood in Fig. 2.

The implementation of FFC and BFE approaches with python is implemented with *RandomForestClassifier* which support in the selection of premier feature in the form of sets. We alter some parameters such as *k_feature=15*; that provided the 15 best average features during the experiments by setting the value 1-25 for our dataset.

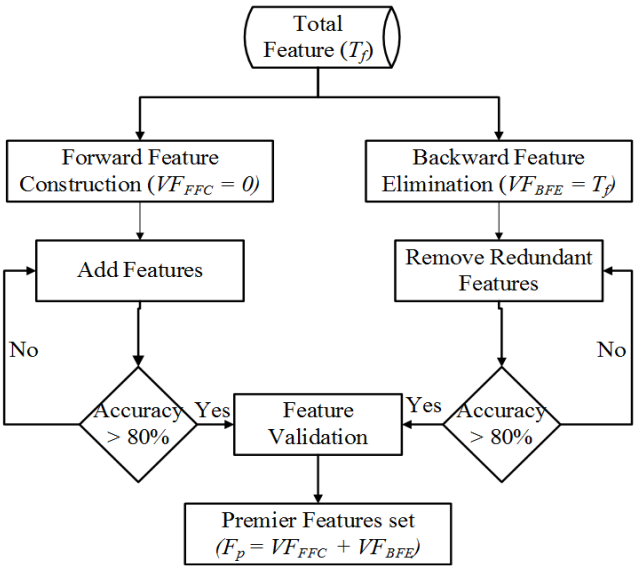

FIGURE 2. The functional criteria of Wrapper Approaches, including FFC and BFE techniques.

A *verbose* will log the progress of feature selection. The scoring parameter *(rou_auc)* is a performance evaluation measure. We took *cv=4*; that refers to *kfold* cross-validation, which split the training data into four sets; three sets will be used for training purposes while 1 for validation. Another critical parameter, *n_jobs= - 1,* which shows, execution will consume all cores of the computer processor.

B. GENETIC ALGORITHM BY MODIFIED FITNESS FUNCTION

A genetic algorithm is practiced for the optimal selection of features to enhance the performance of the feature reduction method [45], [46]. The novelty of this work is the customization of Fitness Function (FF) for the reduction of feature size. The proper functionality of GA can be understood in Fig. 3. In this research, the genetic algorithm randomly performs on *chromosomes (ch)* of a given *population (P)* and steadily improve the fitness values. GA chose two chromosomes *(ch_1 & ch_2)* from population ***P*** for the next generation.

$$\boldsymbol{P} = [P_1, P_2, P_3, \ldots, P_{pop\_size}] \quad (1.1)$$

In Eq. 1.1, *pop_size* is the number of chromosomes in the given population ***P***. Each chromosome P_i consists of some genes or variables p_i and p_j.

$$P_i = [p_{i1}, p_{i2}, \ldots, p_{ij}, \ldots, p_{i\,no\_vars}] \quad (1.2)$$

Wherein i = 1, 2, … *pop_size*; j = 1, 2, … *no_vars,* which denoted the number of variables or genes. The evolution of the population starts from the T to T+1 by the repetition of this procedure. The cumulative selection probability $\hat{Z}_i$ for the chromosome P_i can be defined as,

$$\mathbb{Z}_i = \sum_{j=1}^{i} P_i, \quad i = 1, 2, \ldots pop\_size \quad (1.3)$$

Equation 1.3 shows the selection property of chromosomes. This selection development starts by generating floating-point number, D $\in [0, \mathbb{R}_+]$ for each document D.

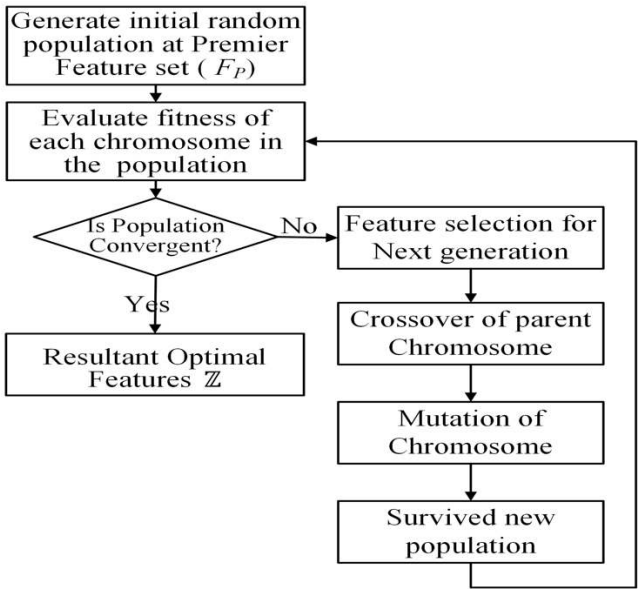

FIGURE 3. An inclusive functionality and relationship of a Genetic algorithm's operators.

In this case, the selection probability is directly proportional to the fitness value, which is based on the Fitness Function (FF) of GA to evaluate the superiority of the proposed solution. For which;

$F_P = \{F_{p1}, F_{p2}, \ldots, F_{pn}\} \triangleq$ the set of *n premier features (Fp)*

$D = \{D_1, D_2, \ldots, D_n\} \triangleq$ the set of documents

$N \triangleq$ the number of documents.

$s(f) \triangleq$ the sparsity of documents using F_P

$x_i = F_i \triangleq$ the value of the ith feature in $\in F_P$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3030642, IEEE Access

tf_{ik} ≜ term frequency of feature $F_i = F_P$ in document $D_k \in D$
df_k ≜ document frequency is the number of the document included $F_k \in F_P$
idf_k ≜ inverse document frequency of feature $F_k \in F_P$ in document

$$D_k \in D = \log\left(\frac{(N-df_k)}{df_k}\right) \quad (2)$$

$$s(f) = 1 - \frac{\sum_{|i=1|}^{|F|}\sum_{|K=1|}^{|D|} f(x_i)}{N^*n - \sum_{|i=1|}^{|F|}\sum_{|K=1|}^{|D|} f(x_i)}, \quad f(x)\begin{cases}1, x = 0\\ 0, x \neq 0\end{cases} \quad (3)$$

Eq. 3 is sparsity ratio which is used to improve the FF's performance. The fitness criteria of chromosome Pi is

$$FF = \frac{\sum_{|i=k|}^{|F|}(idf_k) + \sum_{|K=1|}^{|D|}\min_{i\neq j}(tf_{ik})}{e^{s(f)}} \quad (4)$$

An exponential function $e^{s(f)}$ is used in eq. 4 by utilizing the sparsity ratio to prevent the redundant set of features. The elimination of such features provided significant features with positive values. Eq. 4 represents two terms; the first one calculates the average relevant values of selected features while the second term calculates the average minimum significant value, which have low redundancy value by the chromosome. It provided the highest fitness value when the chromosome of both terms yields a high value. Hence, this modified fitness function can be defined as:

$$FF(p_1, p_2, \ldots p_n) = avg\ (relevance) + avg\ (\min\ (significance)) \quad (5)$$

Eq. 5 is a criterion of FF which signifying the Eq. 4 to consider the chromosome with more fitness value. This FF enabled the GA to compute high relevant and less redundant features that can be vastly converged for optimal features $\mathbb{Z}$.

IV. EXPERIMENTAL IMPEMENTATION AND EVALUATION

This section presents experimental implementation by designing a framework that provides the implementation process of the proposed method and significant results evaluations and discussions.

A. EXPERIMENTAL BASED FRAMEWORK

An integrated framework is designed for its proposed method, which focusing the workflow and the fundamental aspects of hybrid sentiment classification.

Fig. 4 is the simple layout of the proposed framework. The first module of this figure is Data Preparation, which starts from the collection and crawling of the Twitter data. It results in the unstructured tweets. It follows numerous text pre-processing steps to clean the noisy data for meaningful insights and results labeled datasets. The second module is the GAWA Implementation for feature selection, which adopts two Wrapper approaches and a Genetic algorithm with modified fitness function for feature selection and reduction, respectively. The third and last module is the Hybrid Classification Implementation, which applies a lexicon dictionary and four different machine learning algorithms.

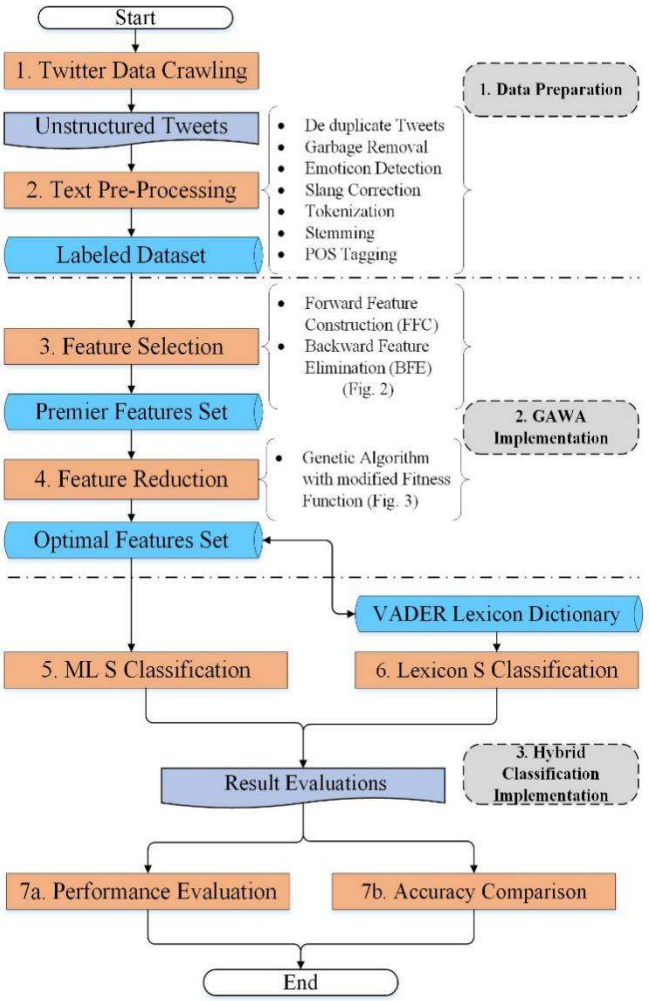

FIGURE 4. A sequential workflow of an experimental based framework for the proposed methodology.

B. DATA PREPARATION

In this research, the data is crawled by Twitter Stream API[1] that is deployed with a python-based crawler program by utilizing the Tweepy library and python3. We use ten different apparel brands (Table 1) names as keywords for the collection of tweets. As a result of this script execution, it received the 66,177 tweets in one week. Due to the ambiguity and complexity of Twitter data for sentiment classification, many researchers [7], [17], [20]–[23], [55] conduct pre-processing techniques before the implementation of their concerned models. Here are some essential pre-processing tasks employed in this paper to make smooth data.

[1] https://developer.twitter.com/en/products/twitter-api

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3030642, IEEE Access

- *De-duplicate the Tweets:* The vital function of De-duplicate is to remove the copied and retweeted tweets.
- *Garbage Removal:* This step removed the non-ASCII character, URLs, hashtags, and web links by using regular expressions.
- *Slang Correction:* It will correct the slang and abbreviated words that are used commonly during conversion due to the character limit of tweet. For example, *"idc" to "I don't care."* It is fixed by predefined dictionaries and translator maps, which convert the slangs and abbreviated words into the original forms.
- *Tokenization:* A process of splitting a text stream into meaningful objects, i.e., words, symbols, or phrases, is called token or tokenization. For which, *LingPipeTokenizer* is used for the token and list of keywords for each document.
- *Stemming:* A common requirement of NLP function but essential for information retrieval methods to reduce various grammatical forms, i.e., noun, adjective, verb, adverb, etc. to its original stem or root form. For example, the word fishing, fisher, and fished to the stem fish. Stemming is adopted by implementing the *Porter Algorithm*, which is the most popular information retrieval stemmer for the English language.
- *POS-Tagging:* The process of tagging a word at its definition and context base on a specific part of speech. A famous tagger from Stanford CoreNLP, *Maxent Tagger,* is utilized for POS tagging.

As a consequence of text pre-processing, the size of resulted in data is reduced due to the removal of noisy tweets. The name of apparel brands, crawled data, and pre-processed data set are shown in Table 1.

TABLE 1
DATA DISTRIBUTION, ACCORDING TO THE NAME OF BRANDS, CRAWLED DATA BEFORE AND AFTER PRE-PROCESSING.

Name of Apparel Brands	Crawled Data	After Preprocessing
Nike	7563	1894
Victoria's Secret	5255	1106
Marc Jacobs	6325	1254
H&M	8659	2491
Gucci	5897	1100
Saint Laurent	4658	1387
Adidas Originals	7961	1879
Calvin Klein	6532	1464
Armani	8657	2861
Toms	4670	1098
Total Data	**66177**	**16534**

C. GAWA IMPLEMENTATION

The core module of this proposed research is feature selection at which the GAWA method is based. As it is discussed, two Wrapper approaches are taking into account for feature selection. The first technique is FFC, which is initiated with zero features. It used the *SequentialFeatureSelector* function from the library of *mlxtend*. The *sklearn* library was downloaded to import the *RandonForestClassifier* and *ROC_AUC* function. The *RandonForestClassifier* is employed to select the optimal parameters as an estimator for the *SequentialFeatureSelector* function. After the creation of the feature selector, a fit function is allowed to pass all training and testing datasets. The implementation process of the second technique (BFE) is the same as FFC. However, there is a little change in the selector parameter to false. Because this process will be revers of FFC, so the parameter attribute is considered as false in BFE.

Table 2 presents the first five and last five IDs of feature sets, which is the output of these two approaches (FFC & BFE). We took 15 features set as estimation; however, some features of some IDs are empty because of unappropriated features. The output of WA contains the premier feature sets of 8243 ID documents.

As a part of feature selection, feature reduction is performed by implementing the customized fitness function of the genetic algorithm. This personalized fitness function is presented in the proposed method (section III). We have employed a library package, including the random function, which helped to initiate the random selection of genes to continue the implementation. After defining the evaluation function, we created a toolbox for all operators, consisting of the fitness function, crossover, and mutation of GA. GA needs more time to extract all optimal features. For the sake of simplicity, the size of the population is stored steady at 100 for all generations of GA. The GA convergence ratio for the required population tremendously depends on the probabilities that will be used in crossover and mutation. The higher proportion of crossover possibility means decreased utilization and elevated exploration; meanwhile, the lower value of this opportunity might also result in insufficient convergence. The typical crossover probabilities' value is between 0.6 and 1.0. Mutation possibility is exceedingly slight as compared to the probability value of crossover. The amount of mutation probability is usually between 0.005 and 0.05 [53]. Table 3 represents the parameters of GA's operators that are used in these experiments.

TABLE 3
THE VALUES AND METHODS OF GA-OPERATORS AND PARAMETERS FOR THE GENETIC ALGORITHM' S-EVOLUTIONS.

GA operators and parameters	Values or Methods
Size of population	100
Parent selection	Random selection
Crossover method	Arithmetic crossover
Crossover probability	0.8
Mutation method	Single point random mutation
Mutation probability	0.05
Population selection	Two best individuals
Maximum number of generations	2000

The resultant individuals from these generations will be evaluated by valid and invalid finesses. As a result, it finds the best individual list which contains the optimal features. Table

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3030642, IEEE Access

4 represents the 3136 optimal features, which are the best set of features as an output of this proposed method (GAWA). In contrast to the other algorithm of feature selection, GAWA succeeded in reducing the 61.95% of features.

D. HYBRID SENTIMENT CLASSIFICATION AND EVALUATIONS

A pre-processed data become capable enough to implement the lexicon dictionaries for polarity detection. In this research, VADER lexical database [27] is employed for the polarity score of all keywords in the document. After the installation of VADER, we imported the *SentimentIntensityAnalyser* class from the *vaderSentiment.vaderSenitment* module. After creating a function to print the sentiment, we defined the *polarity_scores()* function to obtain the score into positive, negative, and neutral. The lexicon-based SA have higher efficiency and accuracy but still is facing the lack of lexical database size, which tried to solve by taking hybrid approaches with multiple ML algorithm. Four different ML algorithms, including NB [33], Sequential Minimal Optimization (SMO), C4.5 – Decision Tree [33], and K-Nearest Neighbor (KNN), are induced to evaluate the accuracy and performance of the proposed method.

1) PERFORMANCE EVALUATION OF ML CLASSIFIERS

The performance evaluation of the machine learning algorithm with the resulted GAWA features is tested with four ML algorithms, which contain C4.5, SMO, NB, and KNN.

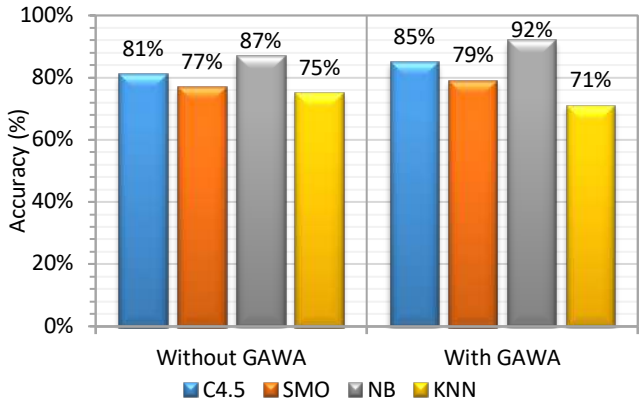

FIGURE 5. Accuracy comparison of machine learning algorithm with and without GAWA features.

Fig. 5 illustrating an overall accuracy of machine learning algorithms that are compared with and without GAWA features. Naïve Bayes performed very well with a maximum accuracy of 92%, which is 5% better than the typical features sets. However, the performances of the rest of the classifiers are also significant. For example, the decision tree algorithm (C4.5) accuracy is 85%. In contrast, the KNN classifier falls behind the SMO classifier with a minimum accuracy level.

2) PERFORMANCE EVALUATION WITH CONFUSION MATRICES

The confusion matrices are applied to validate the classifier's performance and accuracy measures [54], [61]. In this evaluation, Precision, Recall, F-Measure, and Accuracy matrix are adopted with binary classes of positive and negative sentiments. These matrices used the confusion output to find the classes when a prediction is *right [TruePositive (TP), TrueNegative (TN)]* and when the forecast is wrong *[FalseNegative (FN), False Positive (FP)]*.

$$Precision\ (P) = \frac{TP}{TP+FP} \tag{5.1}$$

$$Recall\ (R) = \frac{TP}{TP+FN} \tag{5.2}$$

$$F-measure = \frac{2*(P*R)}{(P+R)} \tag{5.3}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{5.4}$$

A comprehensive evaluation of these matrices (Eq. 5.1, 5.2, 5.3, and 5.4) with different machine learning classifier's performance is clarified in Table 5. We used different sets of optimal features and applied the Precision, Recall, F-Measure, and Accuracy individually for each classifier. It can be concluded that classifiers' performance under the confusion matrices is gradually decreasing with the increasing number of optimal feature sets.

As in real-time, imbalanced class distribution exist in classification problems, which can be tackled by F-Measure due to the utilization of crucial values of False Negative and False Positive. According to the F-Measure, it was observed in Fig. 6 that the mean average value of Naïve Bayes (0.92) for whole feature sets is highest than all other classifiers. Similarly, the accuracy points are also providing the better efficiency of Naïve Bayes and then KNN.

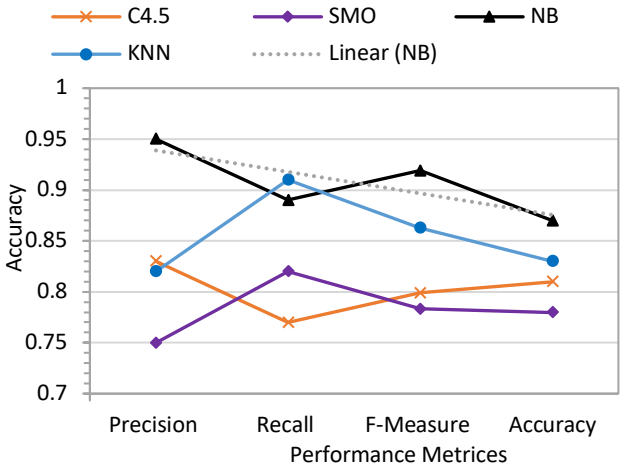

FIGURE 6. Mean performance of ML classifiers under the Confusion Matrices.

In minimum case of features, the Naïve Bayes achieved the best performance; however, its Recall value is less than the C4.5 classifier. Similarly, for the maximum features sets, Naïve Bayes outperform, but its Recall is less than the SMO

classifier. In comparison, the Recall value of the KNN classifier is more excellent than different classifiers; however, SMO performance is lowest than all.

$$NB > KNN > C4.5 > SMO \quad (6)$$

The expression 6 shows the sequence of machine learning classifiers' performances under the combination of Accuracy and F-Measure values with GAWA features. A linear NB bar in Fig. 6 is representing the mean accuracy for all matrices. In conclusion, the Naïve Bayes performance under all matrices is more than other classifiers because Naïve Bayes is more trained for multiple features sets, which was also validated by many studies [56], [57].

3) ACCURACY COMPARISON WITH PREVIOUS WORKS

Comprehensive accuracy comparison of different previous work on sentiment classification for optimal feature selection by various methods, techniques, and the proposed models of feature selection is presented in Table 6. The authors, their applied approach, and achieved the best accuracy in the year sequence is compared with the GAWA approach. It is crystal clear that the proposed GAWA succeeded in making the 92% accuracy with its proposed method, which is more effective and significant than the previous related works.

4) ACCURACY COMPARISON WITH PCA AND PSO

Principal Component Analysis (PCA) [21], [59], and Particle Swarm Optimization (PSO) [22], [50], [60] are unsupervised and supervised learning algorithms, respectively. These algorithms are utilized to reduce the data dimensions for feature selection. GAWA method is based on the same purpose as PCA & PSO. The underlying problem was to achieve the maximum accuracy with the minimum feature sets. Fig. 7 is an accuracy comparison of the proposed method with two existing well-known approaches of feature reduction, PCA & PSO. It shows that the GAWA based approach has an average 84% accuracy with all concerned classifiers while PCA and PSO have an average of 74% and 76% accuracy, respectively. The Naïve Bayes classifier outperforms than other classifiers with all approaches of feature reduction. Meanwhile, PCA with the SMO classifier has minimum accuracy (65%) as well as minimum average accuracy (74%).

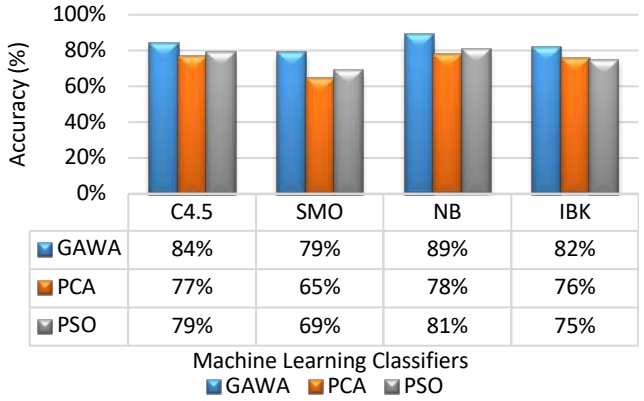

FIGURE 7. Accuracy performance of GAWA with PCA & PSO.

In the evaluation of Fig. 7, it is concluded that the GAWA bases technique provides 11% better accuracy than PCA and 8% than PSO. It proved that the proposed method is more effective for feature selection than the existing and conventional ones.

IV. CONCLUSION AND FUTURE WORK

Research communities are employing various methods and approaches by deploying machine learning techniques for the detection of worthy and precious text features. In this research, one of the biggest challenges of accuracy regarding the massive volume of features is tackled. A novel method (named as GAWA) is proposed for the optimal feature selection. It is based on two Wrapper approaches for premier feature selection, and the Genetic algorithm by its modified fitness function for feature reductions. The GAWA is supported by a primitive framework of hybrid sentiment classification. Four different ML algorithms are performed at the given feature sets of GAWA for the hybrid sentiment classification. The implementation of the Wrapper approaches at Twitter data enabled us to get 8243 premier feature sets, which are reduced by the Genetic algorithm up to 3137 optimal features.

The results show that GAWA succeeded in reducing the feature set up to 61.95%. The performance of resulting features was analyzed with ML classifiers and found that Naïve Bayes outperformed with 92% accuracy than others. The accuracy comparison with five previous works proves the effectiveness of the proposed GAWA method. Furthermore, another accuracy comparison illustrates that the GAWA technique contributes 11% better accuracy than PCA and 8% than PSO.

In future works, this proposed algorithm will be examined with multiple datasets from various sources to select the best features with various categories of syntactic and stylistic features.

REFERENCES

[1] M. Sohrabi, K., and F. Hemmatian, "An efficient pre-processing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study," Multimed Tools Appl 78, 24863–24882 (2019). https://doi.org/10.1007/s11042-019-7586-4.

[2] B. Bong-Hyun, and Il-Kyu Ha. "Comparison of Sentiment Analysis from Large Twitter Datasets by Naïve Bayes and Natural Language Processing Methods," Journal of Information and Communication Convergence Engineering 17, no. 4 (December 31, 2019): 239–45, https://doi.org/10.6109/jicce.2019.17.4.239.

[3] P. Zikopoulos, K. Parasuraman, T. Deutsch, J. Giles, and Corrigan, D. (2012), "Harness the Power of Big Data the IBM Big Data Platform," McGraw Hill Professional, New York, NY.

[4] M. Juncal-Martínez, and Milagros, "Creating emoji lexica from unsupervised sentiment analysis of their descriptions," Expert Systems with Applications 103 (2018) 74–91, https://doi.org/10.1016/j.eswa.2018.02.043.

[5] Alharbi, M. Ahmed Sulaiman and Elise de Doncker. "Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information," Cognitive Systems Research 54 (2019): 50-61, https://doi.org/10.1016/j.cogsys.2018.10.001.

[6] W. Medhat, A. Hassan, and H. Korashy "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J. (2014), http://dx.doi.org/10.1016/j.asej.2014.04.011.

[7] B. Pang, and L. Lee, "Opinion mining and sentiment analysis. Foundations and Trends in information retrieval," (2008) 2, 1-135, https://doi.org/10.1561/1500000001,http://www.nowpublishers.com/article/Details/INR-001.

[8] X. Xie, S. Ge, and F. Hu, "An improved algorithm for sentiment analysis based on maximum entropy," Soft Comput (2019) 23:599–611 https://doi.org/10.1007/s00500-017-2904-0.

[9] P. Kalaivani, and K. L. Shunmuganathan, "Feature Reduction Based on genetic Algorithm and Hybrid Model for Opinion Mining," Scientific Programming, Volume 2015, Article ID 961454, 15 pages, http://dx.doi.org/10.1155/2015/961454.

[10] R. Xia, and Z. Chengqing, "Ensemble of feature sets and classification algorithms for sentiment classification," Inf. Sci. 181, (March 2011), 1138–1152. DOI: https://doi.org/10.1016/j.ins.2010.11.023.

[11] A. Giachanou, and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," ACM Computing Surveys, 49, 1–41, https://doi.org/10.1145/2938640.

[12] N. Zainuddin, and A. Selamat, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," Appl Intell (2018) 48:1218–1232, https://doi.org/10.1007/s10489-017-1098-6.

[13] F. Zablith, "Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models," Applied Mathematical Modelling, Vol 71, pages 569-583, https://doi.org/10.1016/j.apm.2019.02.032.

[14] A. Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches," in IEEE Access, vol. 8, pp. 17722-17733, 2020, https://doi.org/10.1109/ACCESS.2019.2958702.

[15] R. Taleqani, "Public Opinion on Dockless Bike Sharing: A Machine Learning Approach," Transportation Research Record, 2673(4), 195–204. 2019, https://doi.org/10.1177/0361198119838982.

[16] A. Mohammad, R. Hassonah, Al-Sayyed, A. Rodan, M. Al-Zoubi, and I. Aljarah, "An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter," Knowledge-Based Systems, Volume: 192 Article, Number: 105353, https://doi.org/10.1016/j.knosys.2019.105353.

[17] R. K. Thakur, and M. V. Deshpande, "Optimized-Support Vector Machine and Mapreduce framework for sentiment classification of train reviews," Sādhanā 44, 6 (2019). https://doi.org/10.1007/s12046-018-0980-1.

[18] G. Paltoglou, and A. Giachanou, "Opinion retrieval: Searching for opinions in social media," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8830, 193, (2014) https://doi.org/10.1007/978-3-319-12511-4_10.

[19] E. Cambria, B. B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intelligent Systems, volum 28, pp. 15–21, (2013), https://doi.org/10.1109/MIS.2013.30, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6468032.

[20] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," Security Informatics, volum 4, page 9, (2015), https://doi.org/10.1186/s13388-015-0024-xhttp://www.security-informatics.com/content/4/1/9.

[21] C. M. FUNG, and Iqbal, "A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction," (2019), DOI:10.1109/ACCESS.2019.2892852,http://www.ieee.org/publicationsstandards/publications/rights/index.html.

[22] C. A. Qiu, "novel multi-swarm particle swarm optimization for feature selection," Genet Program Evolvable Mach 20, 503–529, (2019). https://doi.org/10.1007/s10710-019-09358-0.

[23] H. Sung-Sam, K. Dong-Wook, and H. Myung-Mook, "Feature selection algorithm based on genetic algorithm using unstructured data for identification for attack mail," Journal of Korean Society for Internet Information 20, no. 1, February 28, 2019, 1–10, https://doi.org/10.7472/jksii.2019.20.1.01.

[24] Z. X. Wang, and Z.P. Lin, "Optimal Feature Selection for Learning-Based Algorithms for Sentiment Classification," Cognitive Computation, 2020. **12**(1): p. 238-248.

[25] A. Mohammad, R. Hassonah, Rizik Al-Sayyed, A. Rodan, and Ibrahim Aljarah, "An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter," Knowledge Based Systems Volume: 192 article number 105353, https://doi.org/10.1016/j.knosys.2019.105353.

[26] J. Wiebe, R. Bruce, M. Martin, T. Wilson, and M. Bell, "Learning subjective language, Computational Linguistics," vol. 30, no. 3, pp. 277–308, 2004, https://doi.org/10.1162/0891201041850885.

[27] M. Rakibul Islam and F. Minhaz Zibran, "A comparison of dictionary building methods for sentiment analysis in software engineering text," In Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '17). IEEE Press, 478–479, 2017, DOI: https://doi.org/10.1109/ESEM.2017.67.

[28] J. Kumar, J. K. Rout, A. katiyar, and S. K. Jena, "Sentiment Analysis Using Weight Model Based on SentiWordNet 3.0.," In: Sa P., Bakshi S., Hatzilygeroudis I., Sahoo M. (eds) Recent Findings in Intelligent Computing Techniques. Advances in Intelligent Systems and Computing, volum 709, Springer, 2018, Singapore, https://doi.org/10.1007/978-981-10-8633-5_14.

[29] A. D. Souza and R. D Souza, "Affective Interaction based Hybrid Approach for Emotion Detection using Machine Learning," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019, pp. 337-342, 10.1109/ICSSIT46314.2019.8987893.

[30] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Approaches to Multi-Objective Feature Selection: A Systematic Literature Review," Ieee Access, 2020. 8: p. 125076-125096.

[31] J. Zhao, and J. M. Liang, "Accelerating information entropy-based feature selection using rough set theory with classified nested equivalence classes," Pattern Recognition, 2020. 107: p. 24.

[32] J. Xiang, W. Yang, and Silamu·Wushouer, "Text sentiment classification algorithm based on feature selection and deep belief network," Journal of Computer Applications, 2019. **39**(7): p. 1942-1947.

[33] A. Madasu, and S. Elango, "Efficient feature selection techniques for sentiment analysis," Multimedia Tools and Applications, 2020. **79**(9-10): p. 6313-6335.

[34] A. Kumar, and A. Jaiswal, "Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter," Multimedia Tools and Applications, 2019. **78**(20): p. 29529-29553.

[35] S. Sharma, and A. Jain, "Hybrid Ensemble Learning With Feature Selection for Sentiment Classification in Social Media," International Journal of Information Retrieval Research (IJIRR), 2020, 10(2), 40-58. doi:10.4018/IJIRR.2020040103.

[36] S. Liu, Y. Liu, F. Wu and W. Fan, "Feature Data Selection for Improving the Performance of Entity Similarity Searches in the Internet of Things," in IEEE Access, vol. 7, pp. 49938-49944, 2019, doi:10.1109/ACCESS.2019.2910736.

[37] M. Zhang, J. Wang, and W. Wang, "HeteRank: A general similarity measure in heterogeneous information networks by integrating multi-type relationships," Information Sciences, 2018. 453: p. 389-407.

[38] Y. Zhou, H. Cheng, and J.X. Yu, "Graph clustering based on structural/attribute similarities," Proc. VLDB Endow., 2009. 2(1): p. 718–729.

[39] F. Thabtah, "Least Loss: A simplified filter method for feature selection," Information Sciences, 2020. **534**: p. 1-15.

[40] O. Gokalp, E. Tasci, and A. Ugur, "A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification," Expert Systems with Applications, 2020. 146: p. 10.

[41] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," in Feature Selection in Data Mining, 2010, pp. 4–13.

[42] Q. Al-Tashi, S. J. Abdulkadir, and H. M. Rais, "Binary Multi-Objective Grey Wolf Optimizer for Feature Selection in Classification," Ieee Access, 2020. 8: p. 106247-106263.

[43] Q. Al-Tashi, H. Rais, S. Jadid, "Feature Selection Method Based on Grey Wolf Optimization for Coronary Artery Disease Classification," Recent Trends in Data Science and Soft Computing, IRICT 2018. Advances in Intelligent Systems and Computing, vol 843. Springer, Cham. https://doi.org/10.1007/978-3-319-99007-1_25.

[44] S. R. Ahmad, A. Abu Bakar, and M.R. Yaaku, "Ant colony optimization for text feature selection in sentiment analysis," Intelligent Data Analysis, 2019. 23(1): p. 133-158.

[45] K. Asha Gowda, A. S. Manjunath and M. A. Jayaram, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning," International Journal of Computer Applications (0975-8887), Volume 1 – No. 7, 2010, https://www.researchgate.net/publication/43656124, DOI:10.5120/169-295.

[46] M. Hammami, S. Bechikh, and C. Hung, "A Multi-objective hybrid filter-wrapper evolutionary approach for feature selection," Memetic Comp. 11, 193–208 (2019). https://doi.org/10.1007/s12293-018-0269-2.

[47] T. Zhou, H. Lu, Y. Wang, and H. Shi, "Collecting genetic algorithm-based PET/CT variable-precision rough high-dimensional feature selecting method, involves changing classification error rate to achieve fitness functions, and obtaining recognition precision for fitness function,". PN CN107679368-AAE UNIV NINGXIA MEDICAL Z9 0UT DIIDW: 2018143772EREF.

[48] B. Chakraborty, "Genetic algorithm with fuzzy fitness function for feature selection," Isie 2002: Proceedings of the 2002 Ieee International Symposium on Industrial Electronics, Vols 1-4. 2002, New York: Ieee. 315-319.

[49] M. Govindarajan, "Sentiment analysis of movie reviews using hybrid method of Bayes and genetic algorithm," International Journal of Advanced Computer Research, (ISSN (print): 2249-7277 ISSN (online): 2277-7970), Volume-3 Number-4 Issue-13 December-2013.

[50] J. Jona, and N. Nagaveni, "A hybrid swarm optimization approach for feature set reduction in digital mammograms," WSEAS Trans. Inf. Sci. Appl., vol. 9, pp. 340–349, 2012.

[51] Y. Wenzhu, and Z. Liang, "An improved genetic algorithm for optimal feature subset selection from multi-character feature set," Expert Systems with Applications Volume 38, Issue 3, March 2011, Pages 2733-2740, https://doi.org/10.1016/j.eswa.2010.08.063.

[52] D. Wang, L. Xie, X. Simon, and Yang, "Support Vector Machine Optimized by Genetic Algorithm for Data Analysis of Near-Infrared Spectroscopy Sensors," Sensors 2018, 18, 3222; Published: 25 September 2018, https://doi.org/10.3390/s18103222.

[53] A. Chatterjee, and M. Rong, "Efficiency Analysis of Genetic Algorithm and Genetic Programming in Data Mining and Image Processing," In I. Management Association (Ed.), Computer Vision: Concepts, Methodologies, Tools, and Applications (pp. 246-272). (2018). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-5204-8.ch010.

[54] S. Zobeidi, N. Marjan, and A. Seyyed Enayatallah, "Opinion mining in Persian language using a hybrid feature extraction approach based on convolutional neural network," Multimedia Tools and Applications 78 (2019): 32357 – 32378, DOI: https://doi.org/10.1007/s11042-019-07993-4.

[55] A. Rasool, R. Tao, M. Kamyab, & T. Naveed, "Twitter Sentiment Analysis: A Case Study for Apparel Brands," Journal of Physics: Conference Series, Vol. 1176, Issue 2, (2019), DOI: 10.1088/1742-6596/1176/2/022015.

[56] F. Nadia Felix Da Silva, F.S. Luiz Coletta, and R. Eduardo Hruschka, "A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning," ACM Comput. Surv. 49, 1, Article 15 (June 2016), 26 pages, DOI: https://doi.org/10.1145/2932708.

[57] S. M. Jimenez Zafra, M. T. Martin Valdivia, E. Martinez Camara and L. A. Urena Lopez, "Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 129-141, 1 Jan.-March 2019, DOI: 10.1109/TAFFC.2017.2693968.

[58] J. S. In, S. Yeong－Wha, and E.D. Jeong, "Verifying the Classification Accuracy for Korea's Standardized Classification System of Research F&E by using LDA(Linear Discriminant Analysis)," Management & Information Systems Review, 2020. 39(1): p. 35-57.

[59] G. Xie, "Method for classifying industrial monitoring data based on incremental principal component analysis (PCA), involves establishing support vector machine (SVM) classification model and verifying model by classifying test set data." Univ Xian Technology (Uyxt-C), Patent Number(s):CN109324595-A; CN109324595-B.

[60] N. Zemmal, "Particle Swarm Optimization Based Swarm Intelligence for Active Learning Improvement: Application on Medical Data Classification," Cognitive Computation: p. 20, to be published.

[61] R. Singh, and V. Goel, "Various Machine Learning Algorithms for Twitter Sentiment Analysis," in Information and Communication Technology for Competitive Strategies, S. Fong, S. Akashe, and P.N. Mahalle, Editors. 2019, Springer International Publishing Ag: Cham. p. 763-772.

A. RASOOL received a BS degree in Software Engineering from the Government College University, Faisalabad, Pakistan, in 2015. Currently, he is doing a Master's degree in Computer Science and Technology from Donghua University, Shanghai, China, through the Chinese Scholarship Council (CSC). He has been an IEEE member 2013-2014 and organized many National Level Technology Seminars, Conferences, Workshops, and Programming Competitions under IEEE. His primary research interests are data mining, machine learning, natural language processing, and social media analysis. He is using machine learning and big data techniques to solve the problems related to huge dimensions of data and social media analysis. He has published four research articles during his Master's degree.

R. TAO received his BS degree in Computer Application and MS degree in Computer Technology from Donghua University, Shanghai, in 1998. He got his Ph.D. in Computer Science and Technology form Donghua University, China. From 1998, he works at Donghua University and has been a Senior Engineer with the School of Computer Science and Technology in 2009. From 2014 to 2015, he was a Visiting Scholar at Old Dominion University, VA, USA. He is the co-author of three books, more than 20 articles, and four patents. His research interests include information systems, cloud computing, and data mining.

M. KAMYAB received a bachelor's degree in computer application from Osmania University, India, in 2015, the Master's degree in computer science and technology from Donghua University, Shanghai, China, in 2019. He is currently pursuing a Ph.D. degree with Donghua University, Shanghai, China. His research interest includes data mining, deep learning, machine learning, natural language processing, social network analysis, and sentiment analysis. He has published four papers in these areas.

H. SHOAIB is doing a Ph.D. in Computer Science and Technology from Donghua University, China. He graduated with a Master's degree in Computer Science and Technology from Donghua University, Shanghai, China, in 2019. He has published three research articles. His research interests include machine learning, artificial intelligence, and network, information security, and privacy protection.

TABLE 2
THE SETS OF PREMIER FEATURES BY THE FFC AND BFE APPROACHES OF WRAPPERS.

ID	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15
0	1.38	11.36	5.11	4.20	6.58	2.08	1.78	0.01	9.52	1.31	16.85	6.96	3.11	12.23	1.77
1										1.29		6.62		10.76	
2	0.00	8.20	7.24	4.54	6.55	1.56	2.47	0.01	7.14	1.58	15.14	6.89	1.90	13.31	1.30
7	2.66	3.04	5.89	1.66	9.77	2.08	1.43	1.25	7.96	1.58	14.42	6.86	5.09	10.40	2.80
10	1.25	11.28	3.74	4.64	8.52	2.30	3.51	0.07	7.61	1.05	15.59	6.27	2.50	11.35	1.35
...															
8239	16526			2.19						2.36		6.54		13.18	
8240	16529			4.20						1.05		6.37		11.29	
8241	16530			3.27						1.31		6.23		12.23	
8242	16532	5.54	5.66	5.86	6.68	9.58	5.31	3.09	3.93	8.79	3.13	13.30	7.63	2.45	14.96	1.51
8243	16533			3.54						1.31		6.14		11.09	

TABLE 4
THE SETS OF OPTIMAL FEATURES ATTAINED BY THE GENETIC ALGORITHM'S IMPLEMENTATION.

ID	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	V15
0	1.32	3.22	6.12	8.31	1.52	1.55	4.88	7.32	1.55	13.01	6.73	10.14	10.99	4.78	6.17
3	1.49	2.10	3.56	7.45	2.14	1.76	2.70	9.91	2.10	14.78	6.90	6.30	12.60	3.78	6.89
4	0.27	9.42	7.23	8.55	2.61	2.52	0.06	9.31	1.29	15.90	6.75	2.94	11.64	1.87	7.01
7	0.94	5.68	5.93	5.96	2.38	2.43	1.53	10.61	1.31	14.89	6.64	4.46	11.64	2.70	6.55
9	1.83	12.20	8.72	11.47	2.49	3.18	0.16	9.03	3.13	15.88	7.35	2.89	12.06	1.74	6.13
...															
3131	1571									1.05		5.97		11.16	
3132	1578									0.77		6.53		10.78	
3133	1580	1.95	4.41	5.83	6.54	2.56	2.12	0.99	6.55	0.81	15.32	6.32	3.52	10.76	1.24	4.12
3134	1581									0.79		6.35		11.87	
3135	1582	2.19	7.58	10.24	12.91	2.18	2.79	0.14	7.83	1.16	15.60	6.47	3.69	10.93	1.60	6.32
3136	1586									0.81		6.26		11.12	

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3030642, IEEE Access

TABLE 5
A COMPARATIVE ANALYSIS OF MACHINE LEARNING CLASSIFIER'S ACCURACY UNDER THE CONFUSION MATRICES WITH THE DIFFERENT NUMBER OF OPTIMAL FEATURES SETS.

Classifier	Performance Matrix	Feature Size				
		10	50	100	200	500
C4.5	Precision	0.88	0.83	0.85	0.81	0.76
	Recall	0.94	0.86	0.84	0.84	0.81
	F-measure	0.90901	0.84473	0.84497	0.824727	0.784204
	Accuracy	0.92147	0.86634	0.86096	0.846281	0.811032
SMO	Precision	0.9	0.85	0.87	0.85	0.79
	Recall	0.87	0.88	0.86	0.84	0.83
	F-measure	0.88474	0.86474	0.86497	0.84497	0.809506
	Accuracy	0.90371	0.86942	0.87426	0.85193	0.821931
NB	Precision	0.96	0.87	0.87	0.82	0.82
	Recall	0.92	0.9	0.89	0.89	0.81
	F-measure	0.93957	0.88474	0.87988	0.853567	0.814969
	Accuracy	0.91235	0.87021	0.84407	0.83171	0.785714
KNN	Precision	0.87	0.82	0.82	0.71	0.69
	Recall	0.89	0.79	0.81	0.75	0.67
	F-measure	0.87988	0.80472	0.81496	0.729452	0.679853
	Accuracy	0.86015	0.81073	0.78301	0.726291	0.646721

TABLE 6
ACCURACY COMPARISON OF THE PREVIOUS RELATED STUDIES WITH THE GAWA FOR THE OPTIMAL FEATURE SELECTION.

Authors & Reference Number	Year	Classifier Algorithm	Applied Approaches	Best accuracy observed (%)
Rui Xia [10]	2010	ME	Joint Part-of-Speech	81.20
Asha Karegowda [45]	2010	Decision Tree C4.5	Wrapper approach in supervised learning	82.71
Nurulhuda Zainuddin [12]	2017	SVM + PCA	Aspect base SA with *SentiWordNe*t and POS-Tag	76.55
Lin Xie [52]	2017	PPSO	Maximum Entropy-PPSO model	87.9
C. M. Fung [21]	2019	Genetic Algorithm	The hybrid approach by using GA	77.9
Proposed GAWA	2020	NB + Genetic Algorithm	Wrapper approach with Genetic algorithm	92