A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification

Osman Gokalp, Erdal Tasci*, Aybars Ugur

Ege University, Department of Computer Engineering, Izmir, Turkey

Abstract

In recent years, sentiment analysis is becoming more and more important as the number of digital text resources increases in parallel with the development of information technology. Feature selection is a crucial sub-stage for the sentiment analysis as it can improve the overall predictive performance of a classifier while reducing the dimensionality of a problem. In this study, we propose a novel wrapper feature selection algorithm based on Iterated Greedy (IG) metaheuristic for sentiment classification. We also develop a selection procedure that is based on pre-calculated filter scores for the greedy construction part of the IG algorithm. A comprehensive experimental study is conducted on commonly-used sentiment analysis datasets to assess the performance of the proposed method. The computational results show that the proposed algorithm achieves 96.45% and 90.74% accuracy rates on average by using Multinomial Naïve Bayes classifier for 9 public sentiment and 4 Amazon product reviews datasets, respectively. The results also reveal that our algorithm outperforms state-of-the-art results for the 9 public sentiment datasets. Moreover, the proposed algorithm produces highly competitive results with state-of-the-art feature selection algorithms for 4 Amazon datasets.

Keywords: Sentiment classification, feature selection, iterated greedy, metaheuristic, machine learning

1. Introduction

As the number of digital text documents increases with the effect of rapid development in information technology, text mining is becoming more and more critical in recent years. Sentiment analysis, also known as opinion mining, is one of the main text classification methods and deals with categoriz-

*Corresponding author

Email addresses: osman.gokalp@ege.edu.tr (Osman Gokalp), arif.erdal.tasci@ege.edu.tr (Erdal Tasci), aybars.ugur@ege.edu.tr (Aybars Ugur)

ing sentimental texts into positive or negative labels. Sentiment analysis can be done generally in three granularity: document-level, sentence-level, and aspect-level (Medhat et al., 2014; Wang et al., 2014, 2015). This study uses document-level sentiment analysis to determine the polarity of text opinions in a given document.

Feature selection is a crucial stage for sentiment analysis (Medhat et al., 2014; Xia et al., 2011). It is one of the dimensionality reduction techniques and can be defined as finding a discriminative subset from all features. It is applied before the classification stage to enhance the predictive performance, reduce the memory requirements, and make data visualization more understandable (Guyon and Elisseeff, 2003). Based on objective functions used, feature selection methods can mainly be divided into two categories, namely, filters and wrappers. Filter-based methods evaluate feature subsets according to a given mathematical criterion, whereas wrapper-based methods employ the predictive performance (e.g., accuracy) for the evaluation. Using the appropriate learning model, wrapper-based methods are able to produce more effective results than filter-based methods. For this reason, wrapper techniques are widely preferred for classification.

Although the high predictive performance it can provide, the main drawback of feature selection is that the feature subset search space grows exponentially as the number of features increases. Furthermore, because wrapper-based methods utilize classification models as an evaluation metric, more computational time is required than using filter-based methods. Metaheuristic algorithms have successfully been used to solve hard optimization problems and provide acceptable solutions in a reasonable time (Talbi, 2009). They make use of problem-specific heuristic information and manage the search process in an efficient way without exploring the whole search space. So, they are ideal candidates to be used to overcome the drawbacks of wrapper-based methods. Some of the examples of metaheuristic-based wrapper algorithms for feature selection are as follows: Genetic Algorithm (Ghareb et al., 2016), Ant Colony Optimization (Wan et al., 2016), Particle Swarm Optimization (Moradi and Gholampour, 2016), Differential Evolution (Hancer et al., 2018), Variable Neighborhood Search (García-Torres et al., 2016), and Tabu Search (Mousin et al., 2016).

Iterated Greedy (IG) (Ruiz and Stützle, 2007) is a metaheuristic algorithm that can successfully be used to solve NP-hard optimization problems. It consists of two fundamental operations that are applied consecutively at each iteration, namely destruction and construction. Destruction operation removes some of the solution components randomly, whereas construction operation adds some of the solution components according to a greedy selection heuristic. Because feature selection requires finding the best subset from all available features, IG is a natural candidate to solve this problem since it can explore the search space by removing/adding features at destruction/construction stages of the algorithm.

This study aims to develop an effective feature selection method for sentiment analysis. We focus not only on achieving high-quality results for text classification tasks but also providing dimensionality reduction. For this purpose, we propose a novel wrapper feature selection algorithm based on IG metaheuristic for sentiment classification. Because of its high performance in sentiment classification, Multinomial Naïve Bayes (MNB) is used as a learner algorithm to use selected features by IG. For the greedy selection part of the IG, we have also developed a filter scores based strategy. A comprehensive experimental study is conducted on commonly-used sentiment classification datasets from Whitehead and Yaeger (2009) and Blitzer et al. (2007) to evaluate the performance of the proposed algorithm. The obtained results are compared with the state-of-the-art results of various sentiment analysis algorithms.

The main contributions of our study to the literature are as follows:

- To the best of our knowledge, this is the first study that employs iterated greedy metaheuristic as a wrapper based feature selection algorithm for sentiment classification.
- A greedy selection procedure that benefits from pre-calculated filter-based scores has been proposed.
- Comprehensive experimental results show that the proposed algorithm could outperform state-of-the-art results for sentiment classification based on the 9 common datasets used.

The remaining sections of this paper are summarized as follows. Section 2 outlines feature selection methods and basic working principles of iterated greedy metaheuristic. Then, Section 3 describes the proposed IG based feature selection algorithm in detail. Next, Section 4 presents an experimental framework for the assessment of the proposed algorithm and compares it with state-of-the-art sentiment classification methods. Finally, Section 5 concludes the paper and discusses possible future studies.

2. The background

2.1. Feature selection methods

Feature selection is a dimensionality reduction approach that can be defined as finding a subset of n feature/features from all features set m, where $n \leqslant m$. Feature selection methods are mainly grouped into two categories, namely filters and wrappers. The following subsections provide detailed information about these two techniques.

2.1.1. Filter-based selection

The filter-based objective function evaluates feature subsets by their information content instead of using predictive models. Filter-based measures are easy to use, fast, and they can be generalized for different classifiers.

This subsection briefly explains some of the main filter-based feature selection measures in the literature, namely chi-square, correlation, gain ratio, information gain, ReliefF, and symmetrical uncertainty coefficient.

Chi-square (χ^2 Statistic)
Chi-square measure evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. It is used to measure the lack of independence between t and c (where t is for term and c is for class) and compared to the χ^2 distribution with one degree of freedom (Dey Sarkar et al., 2014). χ^2 measure is defined for text classification as given by Eq. 1 (Dey Sarkar et al., 2014).

$$\chi^2_{(t,c)} = \frac{D * (PN - MQ)^2}{(P+M) * (Q+N) * (P+Q) * (M+N)} \tag{1}$$

Where D is the total number of documents. P is the number of documents of class c containing term t. Q is the number of documents containing t occurring without c. M is the number of documents class c occurring without t. N is the number of documents of other classes without t.

Correlation
Correlation measure evaluates the worth of an attribute by measuring the Pearson correlation coefficient between the feature and the class. This measure represents the strength of the correlation between them. The correlation coefficient has a value between $+1$ and -1, where $+1$ shows a total positive linear correlation, 0 shows no linear correlation, and -1 shows a total negative linear correlation. It is defined in Eq. 2 (Chen and Wasikowski, 2008).

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \overline{x}_i)(y_k - \overline{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \overline{x}_i)^2 . \sum_{k=1}^{m}(y_k - \overline{y}_i)^2}} \tag{2}$$

Where m is the number of data points, x is the attribute, and y is class.

Gain Ratio

Gain ratio, also known as information gain ratio, reduces the bias for stable evaluation (Duch, 2006). It is calculated by dividing the information gain by attribute entropy (the intrinsic information). The equation of gain ratio is given in Eq. 3.

$$GR = \frac{IG(X)}{IntrinsicInfo(X)} \tag{3}$$

where X denotes the attribute.

Information Gain

Information gain evaluates the worth of an attribute by measuring the gain with respect to the class. It depends on the entropy information. Entropy is a measure of the degree of chaos, or randomness in the system. Information gain represents the amount of information after eliminating uncertainty (Ding and Fu, 2018). The equation of information gain is defined in Eq. 4.

$$IG(X, Y) = H(X) - H(X|Y) \tag{4}$$

Where X is the attribute, and Y is class.

ReliefF

ReliefF is a multiclass, supervised and filter-based feature weighting algorithm that can deal with incomplete and noisy data (Robnik-Šikonja and Kononenko, 2003). ReliefF evaluates the worth of an attribute by repeatedly sampling an instance. It also evaluates the worth of an attribute by taking into account the value of the given attribute for k of its nearest instances of the same and different classes.

Symmetrical Uncertainty

The symmetrical uncertainty coefficient is a modification of information gain which reduces the bias towards the multivalued features (Duch, 2006). It evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class. The equation of symmetrical uncertainty is shown in Eq. 5 (Duch, 2006).

$$SU(X, Y) = 2\frac{IG(X, Y)}{H(X) + H(Y)} \in [0, 1]. \tag{5}$$

Where X is the attribute, and Y is class.

2.1.2. Wrapper-based selection

In wrapper-based selection, the objective function evaluates feature subsets with their classification accuracy rate by using a cross-validation technique. This approach provides more accurate solutions than those of a filter-based approach, however, it can be computationally infeasible, slow and classifier dependent.

2.2. Classification

The classification stage assigns an appropriate category to the pattern with respect to labeled data (e.g., by using supervised learning). Diverse classification models are employed, such as Naïve Bayes (NB), Multinomial Naïve Bayes (MNB), and Bayesian Logistic Regression (BLR). These learner models are briefly explained in the following subsections.

Naïve Bayes

Naïve Bayes classifier is a probabilistic and simple classifier that makes strong assumption conditional independency between all variables given the context of the class (McCallum et al., 1998). There are several Naïve Bayes event models, such as Multi-variate Bernoulli Naïve Bayes, the Poisson Naïve Bayes, and Multinomial Naïve Bayes, according to the fact that how documents are composed of the basic units (Schneider, 2005).

For text classification, the probability that a document d_j belongs to a class c is computed by the Bayes theorem as Eq. 6:

$$p(c|d_j) = \frac{p(d_j|c)p(c)}{p(d_j)} \tag{6}$$

Multinomial Naïve Bayes

The multinomial Naïve Bayes model is a generative model that uses word frequency (e.g., the occurrence of word) information by assuming that the lengths of documents are independent of class in documents instead of the multi-variate Bernoulli event model (e.g., binary vector over the space of words) (McCallum et al., 1998). The classifier is very suitable for text classification tasks because it takes advantage of word frequency information.

Bayesian Logistic Regression

Bayesian logistic regression is a classifier which applies a Bayesian approach to logistic regression. The predictor employs a Laplace prior to prevent overfitting and generates sparse classification models for text data (Genkin et al., 2007). This model is at least as powerful as those produced by ridge logistic regression or support vector machine classifiers combined with feature selection (Genkin et al., 2007).

2.3. Iterated greedy metaheuristic

Iterated Greedy (IG) is one of the metaheuristic algorithms and it is used to tackle hard combinatorial optimization problems. Presented by Ruiz and Stützle (2007), IG applies two main phases iteratively to a solution, namely destruction and construction (see Algorithm 1). In the destruction phase, some of the components are removed randomly from the incumbent solution and a partial solution is produced. In the next phase, called construction, the partial solution is restored by adding new components that are selected by a greedy algorithm. After construction is completed, the resulting candidate solution is evaluated by its quality (e.g., fitness value) and it is accepted if the acceptance criterion is satisfied.

Algorithm 1: Iterated greedy metaheuristic

$S \leftarrow$ GenerateInitialSolution() ;
$S \leftarrow$ LocalSearch(S) ; ▷ `optional`
while *termination condition is not satisfied* **do**
    $S_p \leftarrow$ destruction(S) ;
    $S' \leftarrow$ construction(S_p) ;
    $S' \leftarrow$ LocalSearch(S') ; ▷ `optional`
    if S' *is better than* S **then**
        $S \leftarrow S'$
    end
end
return S

The number of solution components that are to be removed, or destruction size, is the only parameter contained by the IG and it has an important effect on the behavior of the algorithm (Stützle and Ruiz, 2018). To be more specific, if the destruction size is large, most of the solution components are replaced with new ones; hence, the diversification behavior is high. On the other hand, if the destruction size is small, a little part of the solution is changed, so the intensification behavior is high.

3. The proposed iterated greedy algorithm for feature selection

In this section, we explain the proposed optimization algorithm based on Iterated Greedy (IG) metaheuristic to solve the feature selection problem. To achieve this, firstly, we define how a feature selection solution in IG is represented. Then, we give the algorithmic description of the proposed method in detail. After that, the greedy selection procedure and the fitness value evaluation are described.

3.1. Representation of solutions

A solution in IG for feature selection can simply be represented by a binary vector $S = (f_1, f_2, \cdots, f_n)$, where n is the feature count and $f_i \in \{0, 1\}$, $i = 1, 2, \cdots, n$. That is, a feature i is selected if $f_i = 1$ and is not selected $f_i = 0$. An example of destruction and construction operations for a given solution is shown in Fig. 1. In the example, the problem contains 10 features and the incumbent solution has 6 features selected. The destruction procedure removes 2 features (f_1 and f_5) by changing their values to 0. Then, in the construction step, 3 features are selected (f_5, f_6, and f_9) by assigning 1 to their values and the candidate solution is obtained. Note that it is allowed that a feature removed in the destruction can be re-selected in the construction. It is also possible that the number of features deselected may be different from the number of features selected. In this way, the algorithm can explore more within the feature subset space and the number of selected features may vary as the algorithm continues.

3.2. The algorithmic description of the proposed IG algorithm

The proposed IG algorithm (Algorithm 2) for wrapper-based feature selection has 4 main parameters: (i) d is the upper limit of the destruction size, (ii) α controls the greediness level of the selection process, (iii) th_{exp} is the threshold value to start the exploration stage, and (iv) d_{exp} is the d value when the algorithm switches to exploration stage.

At the very beginning of the algorithm, the incumbent solution S is initialized with zero values, meaning that no feature is selected. Accordingly, selected feature count, or sfc, is assigned the value of zero, whereas unselected feature count, or ufc, is assigned the value of total feature count. Then, the main loop of the IG algorithm starts and continues until the number of iterations reaches predefined MAX_ITER value.

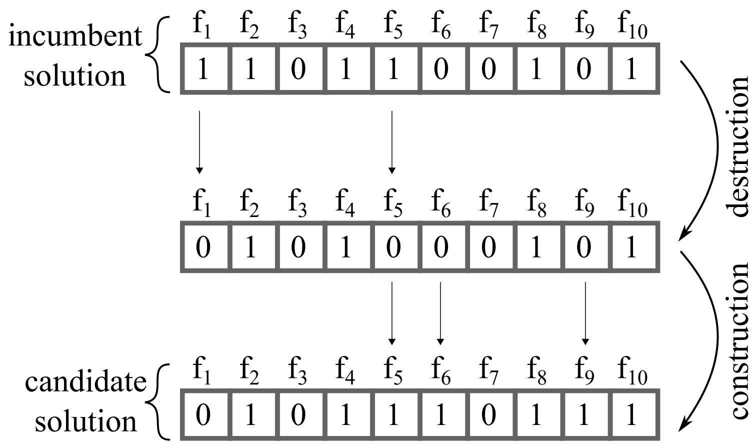

Figure 1: An example of the destruction and construction operations on a 10-feature solution vector.

At each iteration of the main loop, firstly, the copy of the incumbent solution is created as S'. Then, the destruction phase begins and some of the selected features in S' are determined randomly and unselected. After that, in the construction phase, some of the unselected features in S' are retrieved by the greedy selection procedure (see Sect. 3.4) and they are set as selected. Note that the parameter d is not directly used for the destruction size ($destSize$) and the construction size ($constSize$); rather, the random number is selected uniformly up to d. In this way, the number of selected and unselected features can be different at each iteration, so, the selected features count can change adaptively during the algorithm. This behavior makes the algorithm more flexible and helps it to find the required number of features by itself. When the construction is completed, the fitness of the S' is calculated and it is accepted as the new incumbent solution if the acceptance criterion is satisfied (i.e. its fitness is better than the current one).

As a main feature of the proposed algorithm, the exploration stage is additionally used to increase the chance of finding a better solution with a random search. This mechanism has been added to the algorithm because of the observation that no further improvement is obtained after some iterations have passed. Therefore, performing a random search with small modifications (e.g., small d values) can help the algorithm escape from local optima and discover better solutions. To realize the exploration stage, the algorithm checks if the number of iterations reaches th_{exp} value each time before it continues to the next iteration. If this is the case the parameter d and α

are assigned d_{exp} and zero, respectively. The new α value implies that the algorithm performs random search. On the other hand, a small d_{exp} value helps to keep the randomness in balance.

Note that the proposed wrapper-based feature selection algorithm makes use of filter-based greedy selection for the construction phase, so producing new solutions is affected by non-collective information of features (i.e., individual scores per feature). However, this effect is limited because of the randomness in the selection, which is controlled by the greediness level parameter, namely . The diversity in producing solutions, which is provided by both the destruction and construction phases of the IG algorithm, helps explore new feature subsets and supports the collective effect between features. Moreover, the real decision to accept new solutions is dependent on classification accuracy performance, which is calculated by using all the selected features together. In this regard, the acceptance of new solutions is mainly provided by an accuracy score based fitness function that takes the relations between features into account. Indeed, as the new improving solutions are accepted, the good parts of previous solutions are passed to further iterations of IG algorithm and the feature combinations that increase the overall classification accuracy are preserved.

3.3. Evaluation of fitness values

In this study, the fitness function $f(.)$ is evaluated by using the classification accuracy rate measure. The accuracy rate is calculated by the ratio of the total number of correctly classified instances to the total number of instances as in Eq. 7:

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \tag{7}$$

Where TN, TP, FN, and FP denote the number of true negatives, true positives, false negatives, and false positives, respectively. It should also be noted that the 10-fold cross-validation technique is used for every calculation of accuracy.

3.4. A filter scores based greedy selection procedure

The proposed greedy selection procedure selects from unselected features using Greedy Randomized Adaptive Search Procedures (GRASP) (Feo and Resende, 1995) algorithm based on filter-based feature scores. As it is shown in Algorithm 3, firstly, GRASP constructs a candidate list (CL) that includes

Algorithm 2: The proposed wrapper-based iterated greedy algorithm for feature selection

```
input  : d, α, th_exp, d_exp
output: S
filterBasedScores[] ← calculate scores for the features with a
  filter-based method;
for i ← 1 to totalFeatureCount do                    ▷ Initial solution
    S[i] ← 0;
end
sfc ← 0;
ufc ← totalFeatureCount;
for i ← 1 to MAX_ITER do
    S′ ← copy of S;
    d_trunc ← min(d, sfc);
    destSize ← u_rand[0, d_trunc];
    for j ← 1 to destSize do                             ▷ Destruction
        k ← randomly choose a feature from selected features;
        S′[k] ← 0;
        sfc ← sfc − 1;
        ufc ← ufc + 1;
    end
    d_trunc ← min(d, ufc);
    constSize ← u_rand[0, d_trunc];
    for j ← 1 to constSize do                           ▷ Construction
        l ← greedySelection(S′, α, filterBasedScores);
        S′[l] ← 1;
        sfc ← sfc + 1;
        ufc ← ufc − 1;
    end
    f′_S ← computeFitness(S′);
    if f(S′) > f(S) then                        ▷ Acceptance criterion
        S ← S′;
    end
    if i = th_exp then                     ▷ Exploration stage control
        d ← d_exp;
        α = 0.0;
    end
end
```

all the features that have not been selected yet. Then, a restricted candidate list (RCL) is constructed by picking features from CL according to their score values. Finally, a random feature is selected from RCL and returned.

The greediness of the selection process is controlled by the parameter α. Specifically, if $\alpha = 0$, the algorithm becomes completely random selection, i.e. all features are included in RCL, whereas if $\alpha = 1$, the algorithm becomes pure greedy selection, i.e. only the highest-scored feature is included in RCL. Usually, the optimal α value is somewhere between these boundary values and should be set carefully for a given problem.

Algorithm 3: greedySelection

```
input  : S', α, filterBasedScores
output: l
1 CL ← unselected features of S';
2 s_min ← minimum value in filterBasedScores;
3 s_max ← maximum value in filterBasedScores;
4 RCL ← {i ∈ CL | filterBasedScore[i] ≥ s_min + α × (s_max − s_min)};
5 l ← select random feature from RCL;
```

4. Experimental work

4.1. Experimental setup

The proposed algorithm was implemented in Java and all the experiments were performed using WEKA toolkit (Hall et al., 2009) on a computer with the configuration of Intel® Core™ i7 6700 3.40 GHz CPU using a single core.

After some preliminary testing, we adopted the bag-of-words framework, unigram features, term frequency-inverse document frequency (tf-idf) measure to handle sentiment classification datasets. To this end, StringToWordVector was applied to convert string attributes into a set of numeric attributes showing word occurrence information from the text in the strings.

In order to yield performance results for this system, we used three classifiers, namely Naïve Bayes, Multinomial Naïve Bayes, and Bayesian Logistic Regression. The parameter values for the predictive models were set to the default values in WEKA. Besides, we implemented the 10-fold cross-validation for obtaining reliable results. In this process, the dataset is split randomly to the 10 folds. Each fold is used as a test set, while other remaining folds are used as a training set. The process is repeated 10 times and each time different fold is considered for testing. Then, the model is averaged and finalized. After some preliminary testing, the following parameter values were used in the proposed IG algorithm: $d = \sqrt{\frac{\#features}{2}}$, $\alpha = 0.1$, $th_{exp} = \frac{MAX\_ITER}{5}$, and $d_{exp} = 2$.

4.2. Sentiment classification datasets

In this study, in order to evaluate the effectiveness of the proposed algorithm, we utilized 9 public sentiment analysis datasets[1] from Whitehead and Yaeger (2009); Wang et al. (2014) and 4 Amazon product reviews datasets [2] from Blitzer et al. (2007). The names of 9 public datasets are camera, camp, doctor, drug, laptop, lawyer, music, radio and TV, whereas the names of 4 Amazon review datasets are books, dvd, electronics, and kitchen. The 9 public datasets consist of approximately 50% positive and 50% negative reviews. The 4 Amazon product review datasets consist of exactly 50% positive reviews and 50% negative reviews. The characteristics of the datasets are listed in Table 1. The brief descriptions for opinion mining/sentiment mining datasets are shown in Table 2.

4.3. Evaluation measures

To evaluate the performance of the proposed method, we have employed 5 different evaluation measures including classification accuracy rate (ACC), Area Under Curve (AUC), precision, recall, and F-Measure. We also calculated elapsed time for the proposed IG algorithm.

Classification accuracy is computed by dividing the total of true positives and true negatives by the total number of instances. The equation is as shown in Eq. 7.

AUC is the area under the ROC curve for the predictive performance. The area of 1 (e.g., max AUC value) shows a perfect test, the area of 0 (e.g., min AUC value) represents that the predictive model classifies all instances incorrectly.

[1] http://cs.coloradocollege.edu/~mwhitehead/html/opinion_mining.html

[2] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/

Table 1: The characteristics of the used sentiment classification datasets.

Group	**Dataset**	**# of features**	**# of instances**
9 public	Camera	1457	498
	Camp	1810	804
	Doctor	1811	1478
	Drug	1312	802
	Laptop	1840	176
	Lawyer	2123	220
	Music	1441	582
	Radio	1758	1004
	TV	2423	470
Amazon product reviews	Books	28234	2000
	DVD	28310	2000
	Electronics	14943	2000
	Kitchen	12130	2000

Table 2: The descriptions of the used sentiment classification datasets (Whitehead and Yaeger, 2009; Blitzer et al., 2007).

Dataset	**Description**
Camera	Digital camera reviews from Amazon.com
Camp	Summer camp reviews from CampRatingz.com
Doctor	Physician reviews from RateMDs.com
Drug	Pharmaceutical drug reviews from DrugRatingz.com
Laptop	Laptop reviews from Amazon.com
Lawyer	Reviews of lawyers from LawyerRatingz.com
Music	Musical CD reviews from Amazon.com
Radio	Radio show reviews from RadioRatingz.com
Books	Book products reviews from Amazon.com
DVD	DVD products reviews from Amazon.com
Electronics	Electronic products reviews from Amazon.com
Kitchen	Kitchen products reviews from Amazon.com

Precision is the positive predictive value. It is calculated by dividing the number of true positives by the total of true positives and false positives. The equation of precision is defined in Eq. 8.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall is the true positive rate or hit rate. It is calculated by dividing the number of true positives by the total of true positives and false negatives. The equation of recall is defined in Eq. 9.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

F-Measure is the harmonic mean of precision and recall. It is defined in Eq. 10.

$$F - Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}$$

4.4. Preliminary experiments

4.4.1. Performance of various Bayesian-based classifiers

In this section, we analyze and compare the performance of three Bayesian classifiers, namely Bayesian Logistic Regression (BLR), Naïve Bayes (NB), and Multinomial Naïve Bayes (MNB) to use in sentiment classification from the text. So, accuracy rates (10-fold CV) and running times per classifier for 9 sentiment classification datasets without feature selection were collected (see Table 3). According to average accuracy values, MNB could take the highest average accuracy value of 82.20%, while BLR and NB could take 81.71% and 76.29%, respectively. For the running times of the classifiers, MNB is much better than the others, with only 0.06 min. on average.

To statistically analyze the difference between the classifiers, we first performed the Friedman test, which is a non-parametric test for several (more than 2) dependent samples. The Chi-square value of 8.22 and $p = 0.016$ value suggest that there is a statistically significant difference in the accuracy performances of BLR, MNB, and NB.

After finding the significance between classifiers, we further performed a post-hoc test to find the source of this difference. So, we applied the Wilcoxon signed-rank test to each classifier combination. The p values obtained were 0.515, 0.011, and 0.011 for the classifier pairs MNB/BLR, NB/BLR, and NB/MNB, respectively. $p = 0.515 > 0.05$ indicates that there is no statistical

Table 3: The accuracy rates (%) of BLR, MNB, and NB classifiers without feature selection.

Dataset	**BLR**	**MNB**	**NB**
Camera	79.92	82.13	78.71
Camp	87.06	89.18	83.46
Doctor	87.96	89.38	76.59
Drug	71.57	74.44	69.33
Laptop	84.66	83.52	81.82
Lawyer	90.91	87.73	84.09
Music	81.25	78.52	66.67
Radio	75.7	79.38	69.32
TV	76.38	75.53	76.6
Avg. ACC (%)	81.71	**82.20**	76.29
Avg Time (min)	1.54	**0.06**	0.81

significance of the accuracy performances between MNB and BLR. On the other hand, $p = 0.011 < 0.05$ indicates that both MNB and BLR are different from NB.

Overall, we could either choose MNB or BLR as a learner for the sentiment classification in this study because of their higher accuracy values compared to the NB. However, we preferred MNB for the rest of this study, because its running time is overwhelmingly better than that of BLR.

4.4.2. Performance comparison of various filter-based feature methods used in IG

In this section, we analyze and compare the performance of 6 filter methods, namely chi-square, correlation, gain ratio, information gain, ReliefF, and symmetric uncertainty that are used in combination with the proposed IG algorithm. Table 4 reports the average accuracy rates (10-fold CV) of each IG/filter combination for 9 different datasets after running the algorithm through 5,000 iterations. Because IG is a probabilistic algorithm, the reported values are presented as the average of 10 different runs.

To statistically analyze the difference between the algorithms, we first performed the Friedman test. The Chi-square value of 32.262 and $p = 0.000$ value indicates that there is a statistically significant difference in using different filter methods in IG algorithm.

After finding the significance between the effects of filter methods, we

Table 4: The accuracy rates (%) of various filter methods used in IG.

Dataset	**IG with ChiSquare**	**IG with Correlation**	**IG with GainRatio**	**IG with InfoGain**	**IG with ReliefF**	**IG with SymUnc**
camera	96.31	93.25	96.53	96.47	90.48	96.63
camp	97.24	97.28	97.60	97.41	95.01	97.51
doctor	94.17	94.86	95.09	94.42	93.27	94.66
drug	91.20	88.12	91.05	91.20	85.32	91.20
laptop	99.83	97.67	99.83	99.83	97.33	99.83
lawyer	99.41	97.09	99.41	99.41	96.18	99.41
music	92.66	89.69	92.77	92.89	87.10	92.89
radio	91.19	92.42	91.21	91.35	87.32	91.11
tv	95.94	92.94	96.09	96.45	91.57	95.74
Avg.	95.33	93.70	95.51	95.49	91.51	95.44

further performed a post-hoc test to find the source of this difference. So, we applied the Wilcoxon signed-rank test between different combinations of the algorithms. Because there are $\binom{6}{2} = 15$ different combinations, we narrow down search around the IG with gain ratio algorithm as it has the highest accuracy value in comparison with the others. The p values obtained were 0.042, 0.013, 0.889, 0.005, and 0.624 for the algorithms IG with chi-square, correlation, gain ratio, information gain, ReliefF, and symmetric uncertainty, respectively. Statistical results showed that the difference between IG with information gain and symmetric uncertainty algorithms are not significant with regard to $p > 0.05$ values obtained. Nevertheless, we preferred gain ratio filter to be used in the proposed IG algorithm for the rest of this paper since it achieved the highest accuracy value.

4.5. Computational results

The detailed computational results of the proposed IG based feature selection algorithm for sentiment classification are given in Table 5 and Table 6 for the 9 public and 4 Amazon product reviews sentiment datasets, respectively. The values in Table 5 and Table 6 are presented as the average of 10 different runs with 15,000 and 30,000 maximum number of iterations, respectively, and they are calculated based on 10-fold CV. The results show that the proposed algorithm could achieve a 96.45% classification accuracy rate on average for 9 public datasets and 90.74% for 4 Amazon product reviews datasets. Besides, AUC, precision, recall and f-measure values on average are above 90% on average, which indicates the effectiveness of the proposed algorithm for sentiment classification tasks.

Table 5: Computational results of the proposed IG algorithm for 9 public sentiment analysis datasets.

Dataset	**Acc**	**AUC**	**Precision**	**Recall**	**F-Measure**
camera	97.15	0.98	0.97	0.97	0.97
camp	97.99	0.99	0.98	0.98	0.98
doctor	95.64	0.98	0.96	0.96	0.96
drug	92.39	0.95	0.92	0.92	0.92
laptop	99.89	1.00	1.00	1.00	1.00
lawyer	99.59	1.00	1.00	1.00	1.00
music	94.97	0.96	0.95	0.95	0.95
radio	93.05	0.96	0.93	0.93	0.93
tv	97.38	0.98	0.97	0.97	0.97
Avg.	96.45	0.98	0.96	0.96	0.96

Table 6: Computational results of the proposed IG algorithm for 4 Amazon product reviews datasets.

Dataset	**Acc**	**AUC**	**Precision**	**Recall**	**F-Measure**
books	88.00	0.93	0.88	0.88	0.88
dvd	90.25	0.94	0.90	0.90	0.90
electronics	91.63	0.95	0.92	0.92	0.92
kitchen	93.09	0.96	0.93	0.93	0.93
Avg.	90.74	0.95	0.91	0.91	0.91

Fig. 2 illustrates the number of selected features by IG algorithm through iterations for 9 public sentiment datasets employed. Results show that the number of selected features increases rapidly at earlier iterations of the algorithm and it continues with relatively small changes. It can also be seen that the algorithm is very successful at reducing the dimensionality of the problem since it can eliminate majority of the available features. Specifically, it selects approximately 13% of the available features while preserving the high classification accuracy rates.

4.6. Comparison with various feature selection strategies

This section compares the classification performance of our proposed algorithm with 6 filter-based methods that were used in the previous sections

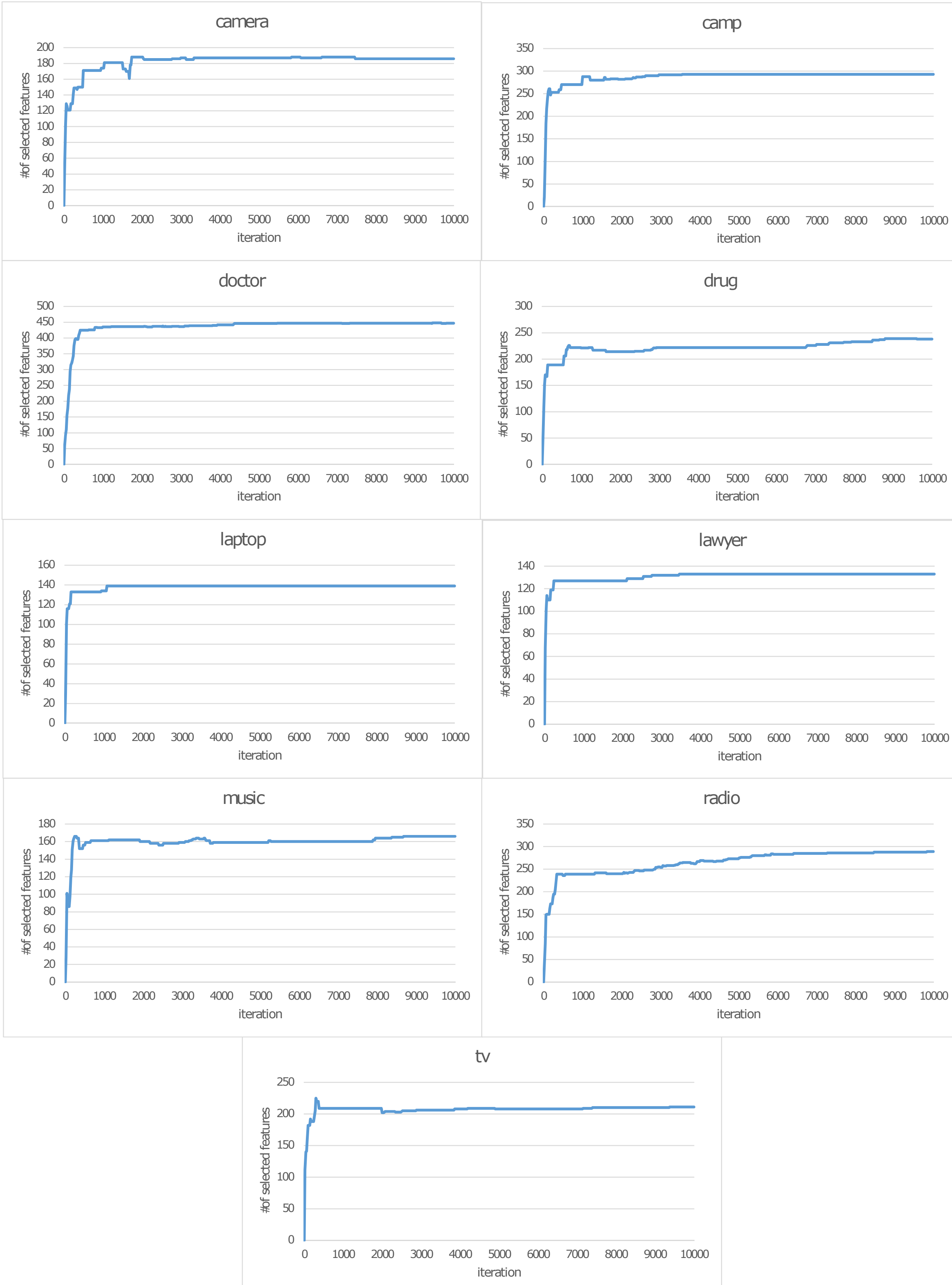

Figure 2: The number of selected features by IG algorithm through iterations for 9 sentiment classification datasets.

and also with the Genetic Algorithm (GA), which is one of the main meta-heuristics and can be used as a wrapper based feature selection.

For the implementation of the GA, the binary-coded chromosome structure that has been mentioned earlier in Section 3.1 was used. Also, evaluations of fitness values were done as it has been described in Section 3.3. In addition, the single-point crossover method was used for producing child solutions and parents were selected using stochastic universal sampling technique. After some preliminary testing, the following parameter configuration was used: mutation rate = 0.01, crossover rate = 1.0, population size = 10, selected features ratio in chromosomes of the starting population = 0.1, and the number of elite chromosomes = 1.

Fig. 3 illustrates the performances of these feature selection algorithms in terms of classification accuracy. All the values were based on 10-fold CV and obtained using MNB as a learner algorithm. For those algorithms that contain randomness in their workings, namely IG and GA, results were obtained by averaging 10 independent runs per dataset. IG algorithm was run through 15000 iterations while GA was run through 1500 iterations. So, both algorithms were run through the same number of fitness function evaluation count as GA does 10 evaluations per iteration. The results show that both proposed IG feature selection and GA feature selection algorithms outperform filter-based approaches as expected. It is also seen that IG clearly outperforms GA as it could achieve higher accuracy values for all the datasets used.

4.7. Comparison with state-of-the-art algorithms for sentiment classification

In this section of the experimental analysis, we compare the performance of the proposed IG based feature selection algorithm with other state-of-the-art sentiment classification algorithms that have used the 9 public datasets.

The selected state-of-the-art algorithms are: a hybrid ensemble pruning approach based algorithm (HEP) of Onan et al. (2017); multi-objective differential evaluation based weighted voting ensemble algorithm (MODE-WV) of Onan et al. (2016); feature unionization (FU) algorithm of Jalilvand and Salim (2017); random subspace (POS-RS) of Wang et al. (2015); and multiple multi-classifier systems based algorithm (MMC) of Yang et al. (2019).

Table 7 reports the accuracy rates in percentage per algorithm/dataset pair. According to average accuracy values, the proposed IG algorithm outperforms all its competitors with 96.45%. Furthermore, it can achieve the highest accuracy rates for 6 datasets out of 9. The results also indicate

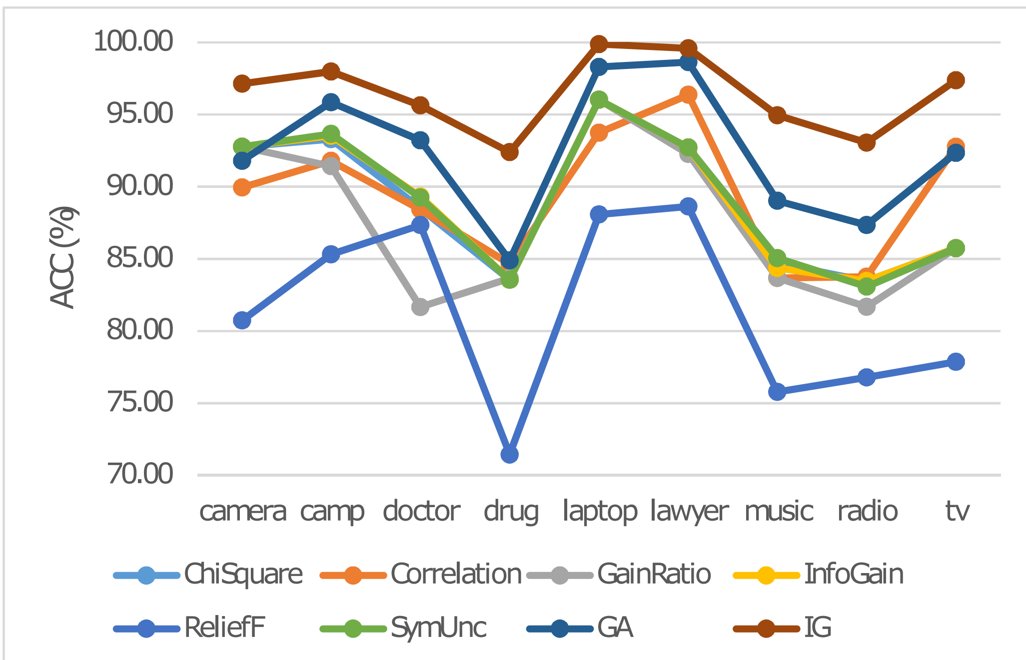

Figure 3: Comparison of various feature selection methods with the proposed IG algorithm.

that the second and the third best-performing algorithms are both ensemble methods that include multiple classifiers.

Table 7: Accuracy rates (%) comparison of the proposed IG algorithm with state-of-the-art algorithms in literature for 9 sentiment datasets. N/A means that no information is provided for a given dataset.

Dataset	IG	HEP	MODE-WV	FU	POS-RS	MMC
camera	**97.15**	95.92	92.87	79.80	76.49	N/A
camp	**97.99**	96.58	93.74	86.00	85.26	82.89
doctor	95.64	**95.65**	91.05	86.10	85.03	83.87
drug	92.39	**94.27**	89.62	69.50	68.82	N/A
laptop	**99.89**	98.92	98.86	78.86	79.79	N/A
lawyer	**99.59**	97.90	97.87	80.91	83.86	N/A
music	**94.97**	94.16	89.82	70.69	69.59	73.18
radio	93.05	**93.37**	88.60	75.30	70.66	67.75
tv	**97.38**	96.73	95.74	79.79	76.06	N/A
Avg.	**96.45**	95.94	93.13	78.55	77.28	-

4.8. Comparison with state-of-the-art feature selection algorithms for sentiment classification

In this section of the experimental analysis, we compare the performance of the proposed IG based feature selection algorithm with other state-of-the-art feature selection sentiment classification algorithms using the 4 Amazon product review datasets, which are mostly employed by these algorithms.

The selected state-of-the-art feature selection algorithms are: A frequency-based integration of different feature subsets (FIFS) feature selection of Yousefpour et al. (2017); recursive feature elimination (RFE) of Ansari et al. (2019); rough set theory (RST) based feature selection of Agarwal and Mittal (2013); feature unionization (FU) dimension reduction of Jalilvand and Salim (2017); and query expansion ranking (QER) dimension reduction of Parlar et al. (2018). Among these studies, the first two algorithms are wrapper-based methods, whereas the remaining algorithms are filter-based methods.

Table 8 reports the accuracy rates in percentage per algorithm/dataset pair. According to the numerical results, the proposed algorithm produced highly competitive results with current state-of-the-art feature selection methods. Specifically, it can achieve the highest accuracy rates for datasets electronics and kitchen with 91.63% and 93.63 %, respectively. In addition, it can produce comparable results for the datasets books and DVD with 88.00% and 90.25%, respectively. Overall, the results indicate the effectiveness of the proposed algorithm for both wrapper and filter-based feature selection in sentiment analysis.

Table 8: Accuracy rates (%) comparison of the proposed IG algorithm with state-of-the-art feature selection algorithms in literature for 4 Amazon product reviews sentiment datasets. N/A means that no information is provided for a given dataset.

Dataset	IG	FIFS	RFE	RST	FU	QER
books	88.00	88.18	90.30	87.70	78.70	**91.60**
dvd	90.25	N/A	88.90	83.20	81.30	**91.70**
electronics	**91.63**	87.03	88.90	83.50	81.80	88.80
kitchen	**93.09**	88.18	N/A	N/A	83.75	91.10

On a conceptual level, we can evaluate that the success of the proposed IG algorithm lies in its local search capability which is highly suitable for subset selection problems. Specifically, each time the destruction and construction operations are applied, new subsets are explored around the solution at hand. The randomness in the destruction phase prevents the algorithm stuck in the local optima, whereas the greedy construction phase guides the algorithm in the huge search space of all possible subsets. Besides, IG is a single-solution based metaheuristic, so, it can produce better solutions than those of population-based wrapper algorithms (e.g., GA and Particle Swarm Optimization) when the lower budget of fitness function evaluations is used.

5. Conclusion

This study proposes a novel wrapper feature selection algorithm based on iterated greedy (IG) metaheuristic for sentiment classification tasks. At each iteration of the IG algorithm, some of the features are removed randomly from the selected features list (destruction) and the new ones are added by filter scores based greedy selection heuristic (construction). As the iterations pass, it is expected that the algorithm improves the best solution found so far.

The preliminary experimental results suggest that the Multinomial Naïve Bayes (MNB) algorithm is very suitable for sentiment classification compared to the other Bayesian approaches. The results also suggest that the proposed algorithm performs well when gain ratio score values are used by a greedy selection heuristic in the construction step. Therefore the proposed algorithm uses MNB as a classifier and gain ratio filter scores as heuristic information for greedy selection.

The performance of the proposed algorithm was tested on both 9 public sentiment and 4 Amazon product reviews datasets that are common in the literature. The results indicate that our algorithm outperforms traditional filter-based feature selection methods as well as GA based feature selection algorithm. The comparison with other algorithms in the literature reveals that our method outperforms state-of-the-art results for the 9 public sentiment datasets. Furthermore, the proposed algorithm produces highly competitive results with state-of-the-art filter and wrapper-based feature selection methods for 4 Amazon datasets.

A natural progression of this study is to apply IG based feature selection algorithm for classification tasks other than sentiment analysis. In addition, the performance of the IG algorithm can further be improved by using ensemble techniques in both classification and greedy selection heuristic stages. Moreover, the main parameters of IG, namely destruction size (d) and the greedy selection level (α) can be determined adaptively during the algorithm execution by using feedback from the search process.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Credit authorship contribution statement

Osman Gokalp: Methodology, Software, Investigation, Writing - Original Draft. **Erdal Tasci:** Methodology, Software, Investigation, Writing - Original Draft. **Aybars Ugur:** Conceptualization, Methodology, Project Administration, Supervision, Writing - Review & Editing.

References

Agarwal, B., Mittal, N., 2013. Sentiment classification using rough set based hybrid feature selection, in: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 115–119.

Ansari, G., Ahmad, T., Doja, M.N., 2019. Hybrid filter–wrapper feature selection method for sentiment classification. Arabian Journal for Science and Engineering 44, 9191–9208.

Blitzer, J., Dredze, M., Pereira, F., 2007. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification, in: Proceedings of the 45th annual meeting of the association of computational linguistics, pp. 440–447.

Chen, X.w., Wasikowski, M., 2008. Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 124–132.

Dey Sarkar, S., Goswami, S., Agarwal, A., Aktar, J., 2014. A novel feature selection technique for text classification using naive bayes. International scholarly research notices 2014.

Ding, J., Fu, L., 2018. A hybrid feature selection algorithm based on information gain and sequential forward floating search. Journal of Intelligent Computing Volume 9, 93.

Duch, W., 2006. Filter methods, in: Feature Extraction. Springer, pp. 89–117.

Feo, T.A., Resende, M.G.C., 1995. Greedy Randomized Adaptive Search Procedures. J. of Glob. Optim. 6, 109–133. URL: `http://link.springer.com/10.1007/BF01096763`, doi:`10.1007/BF01096763`.

García-Torres, M., Gómez-Vela, F., Melián-Batista, B., Moreno-Vega, J.M., 2016. High-dimensional feature selection via feature grouping: A variable neighborhood search approach. Information Sciences 326, 102–118.

Genkin, A., Lewis, D.D., Madigan, D., 2007. Large-scale bayesian logistic regression for text categorization. Technometrics 49, 291–304.

Ghareb, A.S., Bakar, A.A., Hamdan, A.R., 2016. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications 49, 31–47.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research 3, 1157–1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The weka data mining software: an update. ACM SIGKDD explorations newsletter 11, 10–18.

Hancer, E., Xue, B., Zhang, M., 2018. Differential evolution for filter feature selection based on information theory and feature ranking. Knowledge-Based Systems 140, 103–119.

Jalilvand, A., Salim, N., 2017. Feature unionization: a novel approach for dimension reduction. Applied Soft Computing 52, 1253–1261.

McCallum, A., Nigam, K., et al., 1998. A comparison of event models for naive bayes text classification, in: AAAI-98 workshop on learning for text categorization, Citeseer. pp. 41–48.

Medhat, W., Hassan, A., Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal 5, 1093–1113.

Moradi, P., Gholampour, M., 2016. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Applied Soft Computing 43, 117–130.

Mousin, L., Jourdan, L., Marmion, M.E.K., Dhaenens, C., 2016. Feature selection using tabu search with learning memory: learning tabu search, in: International Conference on Learning and Intelligent Optimization, Springer. pp. 141–156.

Onan, A., Korukoğlu, S., Bulut, H., 2016. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Systems with Applications 62, 1–16.

Onan, A., Korukoğlu, S., Bulut, H., 2017. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. Information Processing & Management 53, 814–833.

Parlar, T., Özel, S.A., Song, F., 2018. Qer: a new feature selection method for sentiment analysis. Human-centric Computing and Information Sciences 8, 10.

Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of relieff and rrelieff. Machine learning 53, 23–69.

Ruiz, R., Stützle, T., 2007. A simple and effective iterated greedy algorithm for the permutation flowshop scheduling problem. Eur. J. of Oper. Res. 177, 2033–2049. URL: https://www.sciencedirect.com/science/article/pii/S0377221705008507, doi:10.1016/J.EJOR.2005.12.009.

Schneider, K.M., 2005. Techniques for improving the performance of naive bayes for text classification, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer. pp. 682–693.

Stützle, T., Ruiz, R., 2018. Iterated Greedy. Technical Report. Technical Report TR/IRIDIA/2018-006, IRIDIA, Université Libre de Bruxelles .

Talbi, E.G., 2009. Metaheuristics: from design to implementation. volume 74. John Wiley & Sons.

Wan, Y., Wang, M., Ye, Z., Lai, X., 2016. A feature selection method based on modified binary coded ant colony optimization algorithm. Applied Soft Computing 49, 248–258.

Wang, G., Sun, J., Ma, J., Xu, K., Gu, J., 2014. Sentiment classification: The contribution of ensemble learning. Decision support systems 57, 77–93.

Wang, G., Zhang, Z., Sun, J., Yang, S., Larson, C.A., 2015. Pos-rs: A random subspace method for sentiment classification based on part-of-speech analysis. Information Processing & Management 51, 458–479.

Whitehead, M., Yaeger, L., 2009. Building a general purpose cross-domain sentiment mining model, in: 2009 WRI World Congress on Computer Science and Information Engineering, IEEE. pp. 472–476.

Xia, R., Zong, C., Li, S., 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences 181, 1138–1152.

Yang, K., Liao, C., Zhang, W., 2019. A sentiment classification model based on multiple multi-classifier systems, in: International Conference on Artificial Intelligence and Security, Springer. pp. 287–298.

Yousefpour, A., Ibrahim, R., Hamed, H.N.A., 2017. Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis. Expert Systems with Applications 75, 80–93.