Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 514 – 530

doi:10.3906/elk-1911-116

Development of majority vote ensemble feature selection algorithm augmented with rank allocation to enhance Turkish text categorization

Emin BORANDAĞ*, **Akın ÖZÇİFT**, **Yeşim KAYGUSUZ**
Department of Software Engineering, Faculty of Technology, Manisa Celal Bayar University, Manisa, Turkey

Received: 26.12.2019 • **Accepted/Published Online:** 17.09.2020 • **Final Version:** 30.03.2021

Abstract: The increase in the number of texts as digital documents from numerous sources such as customer reviews, news, and social media has made text categorization crucial in order to be able to manage the enormous amount of data. The high dimensional nature of these texts requires a preliminary feature selection task to reduce the feature space with a potential increase in the prediction accuracy. In this study, we developed an ensemble feature selection method, namely majority vote rank allocation, was developed for Turkish text categorization purposes. The method uses a majority voting ensemble strategy in combination with a rank allocation approach to combine weak filters such as information gain, symmetric uncertainty, relief, and correlation-based feature selection. Thus, the proposed method measures the quality of the features among all features with the majority votes of the filters and ranking allocation. The feature selection efficacy of the method was tested on two datasets, one from the literature and a newly collected dataset. The effect of the obtained features on the classification prediction performance was evaluated on top of the naive bayes, support vector machine J48, and random forests algorithms. It was empirically observed that the developed method improved the prediction accuracies of the classifiers compared to the mentioned filters. The statistical significance of the experimental results were also validated with the use of a two-way analysis of variance test.

Key words: Hybrid feature selection, new enhance, Turkish text categorization, majority voting, ensemble feature strategy, rank allocation

1. Introduction

In the digital era, the information to be processed is generated progressively from various sources such as customer reviews, news, social media, and innumerable digital documents. More than 80% of this information is stored as text, which makes text categorization (TC) a crucial task to manage enormous amount of data [1]. TC is defined as the automated assignment of a given text into predefined categories according to the content of the document. In other words, the automatic extraction of tags from unstructured text according to a predefined set of categories is defined as TC. Developing TC models for this tagging task is valuable for numerous application domains such as information retrieval, customer review analysis, news classification, spam e-mail filtering, topic detection, author identification, bioinformatics, content management, and web page classification [2, 3]. Automated tagging or the classification of texts requires texts to be firstly represented in a model such as vector space (VS) to be handled by machine learning (ML) algorithms. In VS, which in an algebraic model, each word is evaluated as a feature and the value of a feature is weighted depending on metrics such as terms' frequency or term frequency-inverse document frequency (TF-IDF) [3] . In this representation, modelling of

*Correspondence: emin.borandag@cbu.edu.tr

documents as vectors usually produces a high-dimensional sparse feature set, which unveils so-called the curse of dimensionality problem [4] . This high-dimensional model with irrelevant features causes degradation in classification performance and escalation in the run times of ML methods.

High-dimensional feature models of documents include redundant or irrelevant terms that must be eliminated a preprocessing task before classification as. For this aim, efficient feature selection (FS) and feature extraction strategies are proposed to solve the high-dimension issue with expected enhancement in the classifier accuracy [5]. In particular, FS methods aim to obtain as many relevant features as possible to improve the predictive accuracy of the classifiers while reducing the execution times [6]. On the other hand, feature extraction-based approaches generate new features with the use of projection or combination techniques to reach the same objectives. Singular value decomposition (SVD), independent component analysis (ICA), and linear discriminant analysis (LDA) are some of the feature extraction algorithms used for text categorization [7].

FS methods are classified into three main categories: (i) the filters that obtain the most informative features with the use of statistical measures without any ML tasks involved, (ii) the wrappers that combine search strategies and ML methods to obtain the relevant subset of features, (iii) the hybrid techniques that combine filter methods with wrapper approaches in a gradual scheme [8]. Though it is not defined in the literature as a separate group, there are new approaches so called ensemble feature selections. Ensemble feature selection strategies are borrowed from the ensemble learning theory that constructs a set of classifiers for the same problem and then makes a final prediction by taking a vote of their predictions [9]. In the context of FS, an ensemble feature strategy makes use of various feature weighting schemes and combines their output to obtain the optimal feature subset. In this study, a majority vote ensemble feature selection algorithm enriched with a ranking scheme was proposed to obtain the most valuable features that enhance prediction accuracy of widely used ML algorithms. The efficiency of the proposed FS algorithm was tested on two Turkish datasets in the literature [10, 11] and a newly collected dataset from Turkish journal abstracts. The effectiveness of the obtained feature subsets were evaluated using TC algorithms, namely information gain (IG), correlation feature selection (CFS), relief (REL), and symmetrical uncertainty (SU).

The results showed that the proposed FS algorithm significantly enhanced the prediction accuracies of the classifiers and the ensemble method performed considerably better than the single state of the art filter methods. The main contributions of this study were generation of a Turkish ensemble feature selection algorithm that can be used to improve the prediction accuracies of TC classifiers. The rest of the paper is organized as follows: Section 2 introduces the FS strategy in TC domain, Section 3 presents the proposed ensemble feature selection strategy in detail. In Section 4, the datasets are briefly discussed, while in Section 5 the evaluation of the experimental results from a statistical significance perspective is highlighted. Finally, in Section 6, the study is concluded.

2. Materials and methods

In this section, the recent FS algorithms from a TC point of view were overviewed, and, in particularly, related to the Turkish language were emphasized.

2.1. Feature selection methods for text classification

The TC task is the automatic assignment of documents into one or more predetermined classes. As the data to be analyzed grow enormously, automatic classification methods for written information are crucial. To

automate document classification tasks, various TC algorithms and feature engineering methods have been developed. More formally, let $\mid C \mid$ defines the set of classes $C = \{c_1, ...c_{|C|}\}$ for training document set $D = \{d_1, ...d_{|D|}\}$ and V be the representation of distinct words as $V = \{w_1, ...w_{|V|}\}$ occurring in training documents. TC is defined as the estimation of the true class of a document $\mid d \mid$ from the set of labels $\mid C \mid$. Before a document is processed by classifiers, it should be represented in a way such as VS model so that it can be handled with classifier algorithms. In this context, a document d_i is represented as $d_i = \{w_1 i, w_2 i, ...w_V i\}$. This representation, namely one variable for each word from vocabulary $\mid V \mid$, clearly has a sparse or high dimensional nature that makes feature selection inevitable to eliminate irrelevant features. More clearly, the number of candidate feature size for any document can reach hundreds and thousands, exceeding the number of document samples [12]. The importance of a feature or word is measured by its relevancy or contribution to predict the label of the text. As document vectors contain many irrelevant features [13] that do not contribute to the prediction of the class of the document, in some respect the high-dimensional text classification problem becomes a feature selection task to determine the minimal feature subset.

2.1.1. Filter feature selection algorithms

In this section, FS methods, filters, wrappers, hybrid approaches, and ensemble methods are explained. Filters are FS methods that make use of a statistical scoring metric to measure the relevancy of the features (terms). For a given set $S = \{S_1, S_2, ...S_m\}$ of feature size m, the filter methods calculate a score function $\Theta(S_i)$ according to the contribution of feature $(S_i \in S)$ to solve the text classification task. All of the feature weights are ranked depending on their calculated scores, and the features with a score above the threshold are retained while the others are removed [14]. As no learning model is involved in the FS task, filter methods are computationally inexpensive and fast applicable. Thus, filtering-based FS algorithms are widely used among researchers. Filters are used to obtain the most discriminative or compact key-terms that enhance the classification accuracies while decreasing the processing time [15]. There are various filtering methods that are particularly used for text classification purposes. Some of the most commonly used filtering algorithms from text classification literature are document frequency (DF), term variance (TV), distinguishing feature selector (DFS), information gain (IG), term strength (TS), relief (REL), symmetric uncertainty (SU), Chi-square (CHI), and correlation-based feature selection (CFS) [15]. In this study, we designed an ensemble feature selection algorithm, the details of which are given in Section 2.1.4, and which was designed by combining the IG, CFS , REL and SU feature selection algorithms. The main motivation of this ensemble scheme is that feature independence assumption of univariate filters make them incapable to remove redundant features. Therefore, their computational processing advantages are combined with an enhanced voting mechanism that increases filtering strength to obtain the most valuable features. The filter algoritms, namely IG, CFS, REL and SU, are briefly explained below.

i) Information gain: The algorithm measures the mutual dependence of a feature t_j and a category c_i, and it is calculated with the formula given below [16].

$$MI(C; t_j) = \sum_{i=1}^{k} p(c_i, t_j) \log = \frac{p(c_i|t_j)}{p(c_i)} \tag{1}$$

ii) Correlation based feature selection: CFS evaluates the feature scores according to a heuristic correlation function. The algorithm assumes that irrelevant features are uncorrelated with corresponding class label and it selects the subset of valuable features that contributes to class label detection. The scoring function

given below is used to obtain the efficient subset of features S out of all of the features l :

$$M_s = \frac{l\overline{t_{cf}}}{\sqrt{l + l(l-1)\overline{t_{ff}}}} \tag{2}$$

Where $\overline{t_{cf}}$ is the average correlation between the features and corresponding class labels, and $\overline{t_{ff}}$ is the average correlation between two features [17].

iii) Relief: The algorithm filters the features with a scoring scheme that separates instances from separate class labels. For l number of instances the score of the ith feature out of feature size S_i is calculated with the following Equation (3):

$$S_i = \frac{1}{2}\sum_{k=1}^{l} d(X_{ik} - X_{iMk}) - d(X_{ik} - X_{iHk}) \tag{3}$$

,

where (3), M_k is the values for the ith feature of the most adjacent instances to X_k with the same class label, H_k is the ith feature of the most adjacent instances to X_k with a different class label, for the distance measure $d.(.)$ [18].

iv) Symmetric uncertainty: This information theoretic scoring scheme evaluates the quality of the features using the following equation:

$$SU = 2[\frac{IG(X|Y)}{H(X) + H(Y)}] \tag{4}$$

,

where (4), $IG(X|Y)$ is the IG of independent feature X for given class feature Y. Furthermore, $H(X)$ is the entropy of feature X and $H(Y)$ is the entropy of feature X.

Any function used to measure weight or importance of the feature is the core evaluation metric in FS strategies. In wrapper methods, the feature importance is calculated with a classifier that is used as an evaluation function. This function is used in tandem with a search strategy such as sequential backward, sequential forward, or evolutionary search. The evaluation function, i.e. classifier, selects the most valuable features contributing to its prediction performance [19].

Hybrid FS algorithms benefit from or combine the advantages of both filter and wrapper approaches. In particular, as a first step execution time advantage of filter approaches are used to reduce the high dimension of the problem to some extent and then, as a second step, the effectiveness of the wrapper approaches is used to obtain the best subset of features [20].

The ensemble feature selection methodologies are relatively novel compared to the aforementioned subset selection strategies. There are numerous FS methods in the literature with various evaluation strategies. Every FS algorithm has its strengths and weaknesses as mentioned above. On the other hand, there is no well-defined method to prefer a feature selection method to another. As FS is inevitable for high dimensional problems such as text analyzing, researchers have to evaluate various FS schemes for the sake of obtaining enhanced classification accuracies. The ensemble feature selection method based on the ensemble learning theory provides a solution to this issue. In brief, ensemble learning is based on the assumption that the combination of multiple predictors usually provides better performances compared to that of a single predictor. This theory is successfully employed for regression and classification problems. For the FS problem, it enables users to combine the benefits of

various selection strategies rather than choosing a single feature selection model [21]. The main motivation of this study is the combination of relatively weak but computationally fast filters to obtain the discriminative feature subset. As aforementioned, the filter feature selection methods are somewhat weak in terms of obtaining efficient features to enhance classification accuracy. The ensemble learning strategy, namely majority voting, is defined as an ensemble combination technique that makes decisions based on the majority of the votes of predictors to obtain an enhanced classification performance. More formally, considering X $asetofNexamples$ and C denoting $asetofQclases$. For an algorithm set $S = \{A_1, A_2, ..., A_m\}$ of M classifiers used for voting, the majority voting is defined to obtain an overall class prediction for each instance X by the majority decisions of classifiers in the set. If $C_1 \in C$ denotes the class label of an instance X predicted by a classifier A_1, the combination rule F_k is defined as follows:

$$F_k(c_l) = \begin{cases} 1 & c_l = c_k \\ 0 & c_l \neq c_k \end{cases} \tag{5}$$

,

where (5) C_1 and C_k are the class labels. Furthermore, the number of total votes for a class C_k are defined with Equation (6).

$$T_k = \sum_{i=1}^{M} F_k(C_l) \tag{6}$$

From these set of definitions, the class label of an instance X is predicted to be the class that is voted by the majority of predictors. This majority voting mechanism scheme is given in Equation (7) [22].

$$\begin{gathered} c = S(x) = arg\,max T_k \\ k \in \{1, ...Q\} \end{gathered} \tag{7}$$

Majority voting schemes are sometimes used with a weighting mechanism that can further improve the performance of predictions. In other words, certain classifiers are more latent than the others from a prediction point of view. A weighting mechanism gives more importance to the votes of the latent predictors enhancing overall performance. In the case of the issue in the present study, a set of filters $F = \{f_1, f_2, ..f_n\}$ were used to evaluate the value of each feature on the votes of the selections using a rank based weighting scheme.

Briefly, the proposed algorithm makes use of two combined functions to identify the most relevant features depending on a hybrid weighting scheme. The weight or value of a term was evaluated with (i) the majority votes of four IG, CFS, Relief, SU, and (ii) a ranking based scoring scheme. The overall weight or score of each feature was used to obtain the most valuable terms that enhanced the classification performance among all features. The designed approach eliminated the relative weakness of the filters without losing their computational efficiency of them. The detailed structure of the proposed algorithm is given in Section 3.2.

2.1.2. Feature selection in text categorization

Text categorization problems require various FS schemes to be applied before the main classification step. There are numerous studies in the literature that have used an FS strategy as a preprocessing step for the sake of higher classification accuracy and lower computational cost. In this context, the literature review was limited to only a few recent studies particularly using ensemble feature selection. In their recent study, Parlar et al. developed a new FS method on the information retrieval theory and compared their methods with IG, DF filters

on Turkish sentiment classification [23]. Another recent study proposed a relevance frequency based FS metric for Turkish text classification [24]. In their study, Yelmen et al. established a two-step FS strategy that used IG and genetic search (GS) to obtain the optimum feature subset for sentiment classification [25]. Bahassine et al. used an improved FS strategy for Arabic text analysis and developed an improved version of the CHI filter approach to classify a document of six classes [26].

A novel feature selection-based text categorization was evaluated by [27] with the use of Helmholtz principle borrowed from the image processing theory. Sarac et al. used an ant colony search algorithm to reduce feature dimension and also used the optimal features in web data classification [28]. Another group of studies from literature used an ensemble strategy to obtain a valuable feature subset. For example, an ensemble feature selection strategy was evaluated by Hoque et al., using a mutual information concept to create a feature subset out of various feature subsets [29]. Another ensemble-based feature selection algorithm developed for text categorization purposes made use of a modified FS to create features representing all classes with equal importance [30]. There are numerous text categorization studies in the literature that have made use of a FS strategy in some way. However, ensemble feature selection strategies that combine traditional models are infrequent particularly for Turkish text analysis purposes. The aim of this study was to research effectiveness of the ensemble feature selection approach in Turkish text classification. The proposed method is explained in the following section.

3. Proposed method

In this section, the ensemble feature selection strategy, particularly designed for text categorization purposes, is given in detail. The method consists of three steps: (i) text pre-processing, (ii) ensemble feature selection, and (iii) ML evaluation of the quality of the selected features. These steps are explained in the following subsections.

3.1. Text preprocessing

Text representation in the (VS) model uses words as inputs and thus produces a mathematical model that is appropriate for machine learning algorithms. As aforementioned, this model has naturally high dimensional characteristics. Preprocessing is the first step to reduce the high dimension of the model. In this context, the removal of stop words such as 'a-bir', 'that-o', and stemming can be carried out. The stemming process discards suffixes and generates root forms with the same meaning in the VS model. In this study, a well-known Turkish stemmer, namely Zemberek [31], is sed as the first preprocessing of three text corpora.

3.2. Method

The pseudo code of the proposed ensemble feature selection method is given below, and the details of the algorithm is explained as follows:

The ensemble feature selection algorithm has three main processing blocks, namely A, ,B and C, with two main functions that are used to determine the overall weight or value of a feature.

In block A, the main tasks are to (i) obtain features for all datasets, (ii) select classifier algorithms, (iii) obtain the feature lists of the datasets with the use of the IG, CFS, REL, and SU algorithms and (iv) remove the uncommon features that were not selected by IG, CFS, REL and SU.

In block B of the pseudocode, there are two main tasks: (i) calculation of score of a feature for each dataset is obtained using the majority votes of four filter algorithms, (ii) evaluation of the weight of each feature

based on the ranks (positions) they have been selected, (iii) obtaining an overall feature weight for each feature with combining outcomes of the 'vote' and 'rank' functions. In particular, a majority vote principle is used to decide whether a feature to be indexed in the list. For example, if a term is selected by at least three or more filters, then it is retained in the feature subset, otherwise it is removed. The ensemble nature anticipates multiple feature filters to obtain a more critical filter subset than could be obtained from any of the constituent filters alone. This ensemble filtering approach obtains an enhanced feature discrimination power gained from each of the weak filters alone. The second improvement in the feature score evaluation is the relative ranking of the feature in the list. In other words, if a feature is selected in the list with lower indices than it is decided to be a more valuable feature recompensed with a higher score. The overall weight or importance of a feature that contributes to the classification performance then becomes the combination of these two scores calculated using Equation (8).

$$Weight(Feature)_i = \sum_{i=1}^{count} Rank(Feature)_i \times \sum_{i=1}^{count} Vote(Feature)_i \tag{8}$$

The overall weight of each feature in Equation (8) is calculated with a scheme that combines voting and ranking scores. In block C, after calculating the overall weights of the features, the ML algorithms are used to evaluate the quality of the features in terms of classification performance.

The pipeline of the proposed algorithm and its complete architecture are given in Figure 1.

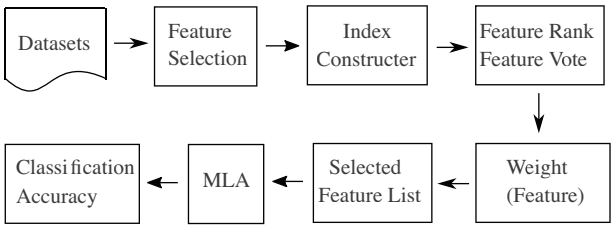


Figure 1. The flow of the designed feature selection strategy and the complete architecture of the algorithm.

4. Datasets

In this study, three corpuses were used a in the evaluation of the designed FS algorithm. As their ML classification performances were already present, TTC- 3600 [10] and Humir [11] were selected from the literature in order to assess the efficiency of the proposed algorithm. In addition, a new dataset, namely U-3000, was collected from Turkish journal abstracts in the engineering, medicine, social sciences and agriculture domains for further assessment of the proposed method [1]. A common characteristics of three datasets was that they had equal distribution of samples in each class and they have no imbalance dataset problem.

4.1. TTC-3600 news dataset

TTC-3600 is a Turkish benchmark news corpus that has six categories each consisting of 600 documents [10]. The brief description of the dataset is given in Table 1.

[1] Ulakbim(2019). TR Dizin Dergi Listesi [online]. Website https://trdizin.gov.tr/statistics/listAcceptedJournals.xhtml [accessed 28 October 2019]

4.2. A new benchmarking dataset

The new dataset was collected in accordance with two aims: (i) enriching the evaluation of the proposed method and (ii) supporting Turkish text analysis research, which has limited benchmark datasets. The new dataset had a collection of 3000 journal abstracts in the engineering, medicine, social sciences and agriculture domains from [1] and was named U-3000 accordingly. Scientific language was used in the dataset. As Turkish is an agglutinative language, the difference in language context caused significant variation in the feature sets of the vector models. Thus, the new dataset was able to further test a wide range of ML algorithms. The brief statistical details of the dataset are given in Table 1. The overall feature (term) size of the dataset was 2590.

Table 1. The description of all datasets

Category	**Number of documents**
Economy	600
Culture-arts	600
Health	600
Politics	600
Sports	600
Technology	600

(a) The description of the TTC-3600 dataset.

Category	**Number of documents**
Engineering	750
Medicine	750
Social Sciences	750
Agriculture	750

(b) The description of the U-3000 dataset.

Category	**Number of documents**
Positive Feedback	5582
Negative Feedback	5582

(c) The description of the Humir-Hotel dataset.

4.3. Humir sentiment dataset

The last corpus used in the evaluation was a hotel review dataset that was researched in [11], which had two sentiment categories, namely positive or negative, and 5582 samples in each category. The details of the dataset are given in Table 1.

5. Experimental study and analysis

In this section, we carry out a set of experiments were did to estimate the performance of the proposed algorithm. We primarily compared the efficiency of our method IG, CFS, REL, and SU with the 'best' results obtained in [10, 11]. As previously mentioned, since the datasets were balanced, thus classification accuracy as the comparison evaluation metric. In addition, the statistical significance of the obtained results were validated with the use of a two-way Analysis of Variance (ANOVA) test. The test was employed to determine whether the empirical results were statistically significant depending on the feature selection methods, classifiers and datasets being independent factors of the analysis.

In the first subsection, details about the evaluation metrics used to analyze the performance of the proposed method were given. As the datasets were balanced accuracy was selected as the reference comparison metric. In subsection 5.2, the details of the classifiers used to evaluate the performance of the produced ensemble feature selection method were given in terms of classification accuracy. The experimental protocol used in the experiments and the obtained results are given in subsections 5.3 and 5.4, respectively.

The experimental results were organized separately for each dataset to clarify the outcomes. The experimental results were organized separately for each dataset to clarify the outcomes. Finally, the statistical significance of the experimental results were validated with the use of a two-way analysis of variance (ANOVA) test. The test was employed to determine whether the empirical results were statistically significant depending on the feature selection methods, classifiers, and datasets being independent factors of the analysis.

5.1. Evaluation metrics

There are many metrics in ML literature to evaluate classifier performances. One of the most frequently used performance metrics for balanced datasets is accuracy. The definition of accuracy is the ratio of the number of correct predictions [32] to the whole number of predictions and is given in Equation (9).

$$A_{CC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

,

where (9), TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

5.2. Classifier algorithms

For the evaluation of the proposed method, detailed experiments were carried out with the four most frequently used classifiers, namely naive bayes (NB), support vector machine (SVM), J48 (java version of C4.5), and random forests (RF) from the text categorization literature. NB is a classifier adopts independence among features and the classifier uses training data to predict the class of a test sample based on the highest posterior probability [33] criteria. Let C denote the class of an instance of $(X_1, X_2, ..., X_m)$ feature vector. While c_j represents jth class label, class of a test instance X is calculated with Bayes' theorem using Equation (10).

$$p(C = c_j | X = x) \alpha p(C = c_j) \prod_{i=1}^{m} p(X_i = x_i | C = c_j) \tag{10}$$

,

where $X = x$ represents the event with condition $X_1 = x_1 \wedge X_2 =, x_2 \wedge ..X_m = x_m$. The equation predicts the class of instance x in terms of the highest probability [33].

SVM makes use of maximization of margin concept to discriminate a set of samples belonging to different classes [33, 34]. For this prediction, an optimal hyperplane is constructed by minimizing an error function $\wedge(w)$ given in Equation (11) iteratively depending on constraints of Equation (12).

$$\wedge(w) = \frac{1}{2} w^T w + c \sum \zeta_i \tag{11}$$

$$\gamma_i = [w^T K(xi) + b] \geq 1 - \zeta_i \; and \; \zeta_i \quad \geq 0, i = 1, ..., n \tag{12}$$

In the equation, w is coefficient vector, b is a constant, and ζ_i denotes the parameters for misclassification. For a training instance, i features for a class label γ_i are given by X_i . In this formulation, K is the kernel function that is used to transform input data into higher feature space in the generation of a nonlinear decision boundary [33, 34].

J48, java version of C 4.5 decision tree algorithm, is widely used in classification and, regression. In a classification problem, the internal, branch, and leaf nodes of the tree denote test on the feature, the outcome of the test, and the class labels, respectively. Furthermore, the paths from root to leaf forms generate the corresponding classification rules. Any node in the tree is basically used as an estimation criterion to determine the relevant features depending on entropy reduction and information gain [35]. The entropy is defined with Equation (13) given below.

$$E = -p^p x \log_2(p^p) - pN \times log_2(pN) \tag{13}$$

,

where p^p is the positive training instance ratio, p^N is the negative training instance ratio. RF makes use of a voting mechanism to predict class of an instance from the overall predictions of decision trees in the forest. The dataset used for training is obtained with bagging, in which samples are arbitrarily drawn with the replacement scheme. A tree in the forest is accomplished with a feature selection criteria and pruning strategy. A widely used FS approach is the Gini index, which is calculated with Equation (14).

$$\sum \sum_{i \neq j} (\frac{f(C_i, T)}{|T|})(\frac{f(C_j, T)}{|T|}) \tag{14}$$

,

where T is the training set, C_i is a class, and $f(C_i, T)/|T|$ is the probability measure for a selected sample belonging to corresponding class C_i. The selected features are used to obtain the best split. Hence, RF with N trees makes the classification with the majority of votes of N trees in the forest [13].

5.3. The experimental protocol

The FS and classification results of the experiments were obtained with the experimental protocol that can be briefly explained as follows: we first have preprocessed the three datasets (firstly the IG, CFS, REL and SU FS methods were applied, and four feature sets were obtained for each of the datasets); then, majority vote rank allocation (MVRA) FS algorithm was applied to the original datasets, and three more feature subsets were obtained. Four classification algorithms, namely J48, RF, NB and SVM, with default parameters were used to inspect the performance of the feature subsets on top of the ten-fold-cross-validation scheme. All classification experiments were repeated 10 times and the average accuracies were calculated.

5.4. Result of the experiments

In this section, the experimental results obtained with the use of the experimental protocol mentioned in the previous section are provided. The obtained results were handled three subsections for each dataset, and then the overall results were interpreted.

5.4.1. Comparison of MVRA with the univariate filters of the TTC-3600 dataset

The experimental results of the TTC-3600 dataset are given in Table 2 below. As can be seen from the table inspected, we can perceive that the proposed ensemble approach, namely MVRA, excelled all of the single basic

filters, namely IG, SU, Relief, and CFS in terms of classification performances. In other words, the collaboration of weak filters generated a better subset of features that resulted in enhanced prediction accuracies. Furthermore, the feature selection performance of the proposed approach was also observed to be significantly higher compared to "all features". In particular, NB had a noticeable response to the ensemble feature selection algorithm. An overall visual summary of Table 2 was generated in Figure 2a.It can be seen clearly from Figure 2a that MVRA had an enhanced average classification performance with significantly better statistical distribution in terms of the minimum and maximum predictions of classifiers. In a study, in which TTC-3600 was investigated, [10] various feature selection strategies were used to obtain the best prediction accuracy of 91.03%. As can be seen from Table 2, the proposed ensemble feature selection algorithm increased the prediction accuracy to 92.44 % which is a meaningful improvement.

5.4.2. Comparison of MVRA with the univariate filters of the U-3000 dataset

The experimental results for the newly collected dataset, namely U-3000, are provided in Table 2. When the table is examined, it can be observed that MVRA had better prediction accuracies, apart from the CFS filtering approach, which had a negligible improvement. While the difference between the CFS and MVRA approaches was minor, the remaining prediction performances verified the FS ability of MVRA to significantly enhanche the prediction abilities of classifiers. Excluding CFS's slight improvement, MVRA performed better in all of the experiments including the "all feature" setup. More precisely, the prediction performance of the algorithms are increased with the ensemble feature selection compared to the single feature selection algorithms. In particular, MVRA has a significant contribution to classification performances of algorithms while compared to the unprocessed versions of datasets, namely with "all features". We summarize this comparison in Figure 2. The overall inspection of Figure 2 demonstrates the efficiency of feature selection particularly for classification performances. The inspection of Figure 2 shows the significant performance of MVRA in feature selection potential in terms of classification accuracy compared to "all features".

5.4.3. Comparison of MVRA with the univariate filters of the Humir-Hotel dataset

One of the benchmarking datasets from the literature was the Humir-Hotel dataset, which is collected for sentiment analysis purposes. The experimental results for MVRA, IG, CFS, Relief, SU, and "all features" are provided in Table 2. The performance of the MVRA seen from Table 2 is that the algorithm still improves the performances of the algorithms except two cases of SU. On the other hand, the contribution of MVRA to the classification performances was acceptable. In particular, when combined with NB and SVM, MVRA may provide satisfactory classification results. The dataset is also studied by B. Ersahin et al. in [11] and the researchers obtained 91.96% classification accuracy as their best value. In terms of classification performance, MVRA based NB and SVM have better accuracies with 92.2% and 92.62%, respectively. The brief visual summary of Table 2 is provided in Figure 2c.

In particular, MVRA significantly contributed to the classification performances of the algorithms when compared to the unprocessed versions of the datasets, namely with all features. This comparison is summarized in Figure 2. The overall inspection of Figure 2 shows the significant performance of the FS potential of MVRA in terms of classification accuracy compared to all features. Though computational cost(time) calculations for the experiments were not included, it is evident that reduced feature dimensions will reduce corresponding evaluation time.

Table 2. The description of all dataset.

Algoritms	All F.	IG	SYM	REL	CO	Best results in [10]	MVRA
J48	78.06	78.72	78.61	77.97	78.11	79.00	**79.07**
NB	82.94	91.22	91.22	76.11	78.22	87.17	**92.44**
RF	88.53	88.22	88.03	87.72	88.30	**91.03**	90.44
SVM	86.03	86.11	86.11	86.94	86.63	86.03	**87.91**

(a)Experimental ACC results for the TTC-3600 dataset.

Algoritms	All F.	IG	SU	REL	CFS	MVRA
J48	78.66	79.56	79.73	78.83	**79.96**	79.56
NB	88.16	89.46	89.20	88.66	89.23	**91.87**
RF	90.93	91.00	91.53	90.13	90.76	**91.57**
SVM	91.2	90.36	90.83	90.26	90.70	**92.4**

(b)Experimental ACC results for the U-3000 dataset.

Algoritms	All F.	IG	SU	REL	CFS	Best results in [11]	MVRA
J48	86.10	88.56	88.69	88.16	88.59	**89.2**	88.63
NB	83.80	83.56	83.56	89.36	91.13	90.9	**92.20**
RF	91.14	91.98	**92.09**	91.00	91.90	Not Exist.	91.00
SVM	92.47	**92.62**	92.41	92.07	**92.62**	92.0	**92.62**

(c)Experimental ACC results for the Humir-Hotel dataset.

Algoritms	All F.	IG	SYM	REL	CO	Best results in [10]	MVRA
J48	78.0	78.7	78.5	77.9	78.0	Not Exist.	**78.9**
NB	82.9	91.2	91.2	76.1	78.7	Not Exist.	**92.1**
RF	88.3	88.1	88.0	87.7	88.2	Not Exist.	**90.2**
SVM	85.8	86.1	86.1	87.0	86.6	Not Exist.	**87.8**

(d)Experimental FM results for TTC-3600 dataset.

Algoritms	All F.	IG	SYM	REL	CO	MVRA
J48	78.6	79.7	79.8	78.9	**80.1**	79.7
NB	88.2	89.5	89.2	88.7	89.2	**91.9**
RF	90.9	91.0	91.5	90.1	90.8	**91.6**
SVM	91.2	90.4	90.8	90.3	90.7	**92.4**

(e)Experimental FM results for U-3000 dataset.

Algorits	All F.	IG	SYM	REL	CO	Best Results in [11]	MVRA
J48	86.0	88.6	88.5	88.2	88.5	**88.9**	88.6
NB	83.8	83.6	83.6	89.4	91.1	89.9	**92.1**
RF	91.1	92.0	**92.1**	91.0	91.9	Not Exist.	91.0
SVM	92.4	**92.6**	92.4	92.1	**92.6**	92.0	**92.6**

(f)Experimental FM results for Humir-Hotel dataset.

5.5. Statistical analysis

In order to measure the statistical significance of the experimental results, a two-way ANOVA test was employed. In this context, the differences among the various groups of the experimental result were found and the statistical

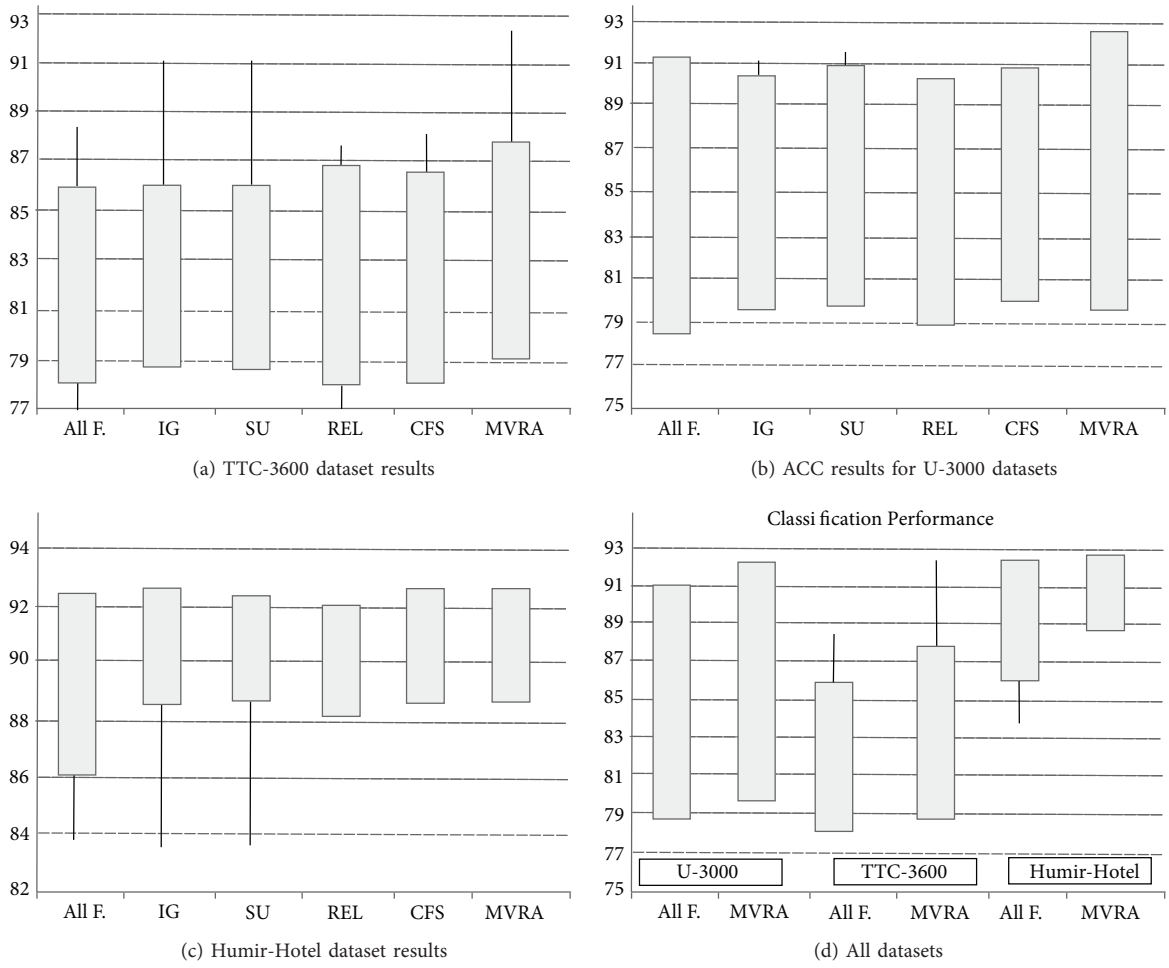


Figure 2. The comparison of ACC performances of MVRA and base feature selection algorithms for all datasets.

significance of these differences were determined depending on independent factors. In the assessment of the ANOVA test, the factors to be analyzed were the FS strategies, machine learning predictors, feature sets, and three text categorization datasets. In the analysis, the responses were evaluated in terms of ACC. All of the mentioned statistical evaluations were implemented with the use of Minitab statistical software suit. The factor information and the corresponding statistical values of the two-way ANOVA test are given in Figure 3. The parameters of the statistical test corresponded to degrees of freedom (DF), adjusted sum of squares (Adj SS), adjusted mean square (Adj MS), F-value, and P-value.

The Adj SS of the squares are measures of variation and furthermore the Adj SS term of the squares denotes the increase in the regression sum. All terms are used for the distribution model of the residuals that are used to inspect the quality of the fit in regression.

The regression line and normal distribution of the residuals shows the empirical validity of the experiments in terms of the ACC metric.

Figure 3a shows that the proposed ensemble feature selection strategy improved classification accuracies with a 95% confidence based on the P-values. Figure 3a shows the empirical comparisons of the three sets in terms of mean ACC, namely the performance of the classifiers, effect of the FS strategies, and the accuracies

for each dataset. The highlights that can be drawn from the Figure 3b are the proposed FS strategy, MVRA, and improved prediction performances. The corresponding analysis is presented in Figure 3c.

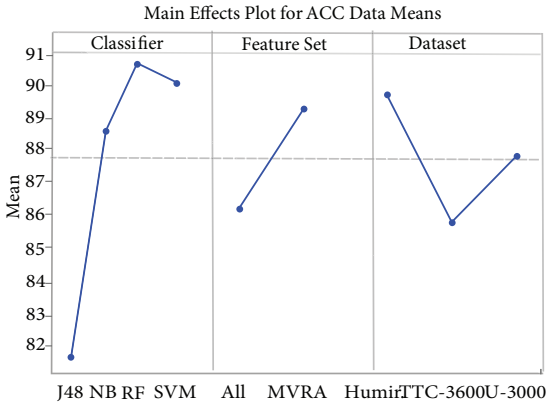
Factor Information

Factor	Type	Levels	Values
Classifier	Fixed	4	J48; NB; RF; SVM
Feature Set	Fixed	2	All; MVRA
Dataset	Fixed	3	Humir-Hotel; TTC-3600; U-3000

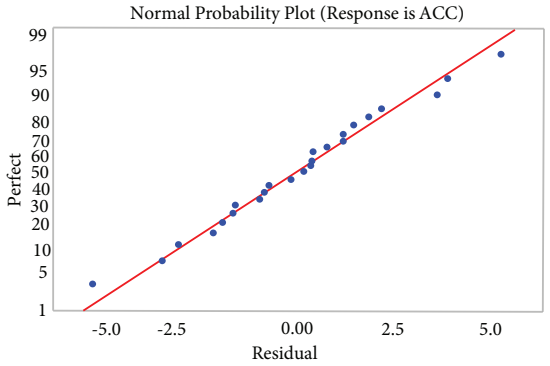
Analysis of Variance

Source	Df	Abj SS	Abj MS	F-Value	P-Value
Classifier	3	315.09	105.030	12.99	0.000
Feature Set	1	45.57	45.568	5.64	0.030
Dataset	2	63.12	31.558	3.90	0.040
Error	17	137.47	8.087		
Total	23	561.25			

(a) Statistics results of all datasets



(b) Main effects plot for ACC



(c) Normal Probability Plot for ACC

Figure 3. ANOVA statistics results of the proposed feature selection strategy.

Conclusion

The selection of discriminative or valuable feature subsets from a high dimensional feature space is one of the primary tasks of the text classification process. The high dimensional nature of texts having a large number of

features causes an increase in computational cost and a decrease in the prediction qualities of the classifiers. The expectation from FS is to decrease computational complexity while increasing prediction accuracy. However, there is no universal FS method serving or guaranteeing a particular accuracy goal. There is an ongoing research in feature engineering to design powerful FS algorithms. In this respect, in the present study, an ensemble learning theory concept, namely voting strategy, was combined with a ranking scheme to obtain an effective FS strategy using simple feature filters. The proposed method, MVRA, was tested on two Turkish text corpus from the literature and a newly collected dataset. The performance of the proposed method was evaluated against 'IG, SU, REL, and CFS based features and all features' on top of J48, NB, RF and SVM classifiers. Remarkable results were obtained in terms of classification accuracies. The overall conclusion of the empirical experiments is that the proposed FS strategy is successful. From these two points, it can be concluded that the MVRA algorithm in tandem with SVM and NB are successful in text categorization tasks particularly for Turkish. In future studies, we plan to extend the use of this approach to other languages to observe its performance.

Data availability

The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments

This work is supported by Manisa Celal Bayar University Departmant of Scientific Research Projects. Project No: 2019-057

Conflicts of Interest

The study was performed as part of the employment of the authors at Manisa Celal Bayar University.

References

[1] Ghareb AS, Bakar AA, Hamdan AR. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications 2016; 49: 31-47. doi : 10.1016/j.eswa.2015.12.004

[2] Meena MJ, Chandran KR. Naïve Bayes text classification with positive features selected by statistical method. In: IEEE 2009 First International Conference on Advanced Computing; Chennai, India; 2009. pp. 28-33. doi: 10.1109/ICADVC.2009.5378273

[3] Labani M, Moradi P, Ahmadizar F, Jalili M. A novel multivariate filter method for feature selection in text classification problems. Engineering Applications of Artificial Intelligence 2018; 70: 25-37. doi: 10.1016/j.engappai.2017.12.014

[4] Altinel B, Ganim MC. Semantic text classification: A survey of past and recent advances. Information Processing and Management 2018; 54 (6): 1129-1153. doi: 10.1016/j.ipm.2018.08.001

[5] Jin C, Ma T, Hou R, Tang M, Tian Y et al. Chi-square statistics feature selection based on term frequency and distribution for text categorization. IETE Journal of Research 2015; 61 (4): 351-362. doi: 10.1080/03772063.2015.1021385

[6] Costa H, Galvao LR, Merschmann LHC, Souza MJF. A VNS algorithm for feature selection in hierarchical classification context. Electronic Notes in Discrete Mathematics 2018; 66: 79-86. doi: 10.1016/j.endm.2018.03.011

[7] Biricik G, Diri B, Sonmez AC. Abstract feature extraction for text classification. Turkish Journal of Electrical Engineering and Computer Science 2012; 20 (Sup.1): 1137-1159. doi: 10.3906/elk-1102-1015

[8] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics. 2015; 2015: 198363. doi: 10.1155/2015/198363

[9] Dietterich TG. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science. Berlin, Heidelberg; 2000. pp. 1-15. doi: 10.1007/3-540-45014-9_1

[10] Kilinc D, Ozcift A, Bozyigit F, Yildirim P, Yucalar F. et al. TTC-3600: A new benchmark dataset for Turkish text categorization. Journal of Information Science 2017; 43 (2): 174–185. doi 10.1177/0165551515620551

[11] Ersahin B, Aktas O, Kilinc D, Ersahin M. A hybrid sentiment analysis method for Turkish. Turkish Journal of Electrical Engineering & Computer Sciences 2019; 27: 1780–1793. doi: 10.3906/elk-1808-189

[12] Novovicova J, Malik A. Information-theoretic feature selection algorithms for text classification. In: IEEE 2005 IEEE International Joint Conference on Neural Networks; Montreal, Quebec, Canada; 2005. pp. 3272-3277.

[13] Chakrabarti S. Mining the Web: Discovering Knowledge from Hypertext Data. Amsterdam: Morgan Kaufmann, 2003.

[14] Deng X, Li Y, Weng, J, Zhang J. Feature selection for text classification: A review. Multimedia Tools and Applications 2019; 78 (3): 3797-3816. doi: 10.1007/s11042-018-6083-5

[15] Gunal S. Hybrid feature selection for text classification. Turkish Journal of Electrical Engineering and Computer Sciences 2012; 20 (Sup.2) : 1296-1131. doi: 10.3906/elk-1101-1064

[16] Meng J, Lin H, Yu Y. A two-stage feature selection method for text categorization. Computers & Mathematics with Applications 2011; 62 (7): 2793-2800. doi: 10.1016/j.camwa.2011.07.045

[17] Czarnowski I, Wosiak A, Zakrzewska D. Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis, Complexity 2018, 2018: 2520706. doi: 10.1155/2018/2520706

[18] Tang J, Alelyani S, Liu H. Feature selection for classification: a review. In: Aggarwal CC, editor. Data Classification: Algorithms and Applications. Boca Raton, FL, USA: CRC Press, 2013, pp. 37-64.

[19] Panthonga R, Srivihokb A. Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. Procedia Computer Science 2015; 72: 162-169. doi: 10.1016/j.procs.2015.12.117

[20] Manbari Z, Tab FA, Salavati C. Hybrid fast unsupervised feature selection for high-dimensional data. Expert Systems With Applications 2019; 124: 97-118. doi: 10.1016/j.eswa.2019.01.016

[21] Bolon-Canedo V, Alonso-Betansoz A. Ensembles for feature selection: a review and future trends. Information Fusion 2019; 52: 1-12. doi: 10.1016/j.inffus.2018.11.008

[22] Bouziane H, Messabih B, Chouarfia A. Profiles and majority voting-based ensemble method for protein secondary structure prediction. Evolutionary Bioinformatics 2011; 7: 171-189. doi: 10.4137/EBO.S7931

[23] Parlar T, Ozel SA, Song F. QER: a new feature selection method for sentiment analysis. Human-centric Computing and Information Sciences 2018; 8 (1): 10. doi: 10.1186/s13673-018-0135-8

[24] Sahin DO, Kilic E. Two new feature selection metrics for text classification. Automatika 2019; 60 (2): 162-171. doi: 10.1080/00051144.2019.1602293

[25] Yelmen I, Zontul M, Kaynar O, Sonmez F. A novel hybrid approach for sentiment classification of Turkish Tweets for GSM operators. International Journal Of Circuits, Systems And Signal Processing 2018; 12: 637-645.

[26] Bahassine S, Madani A, Al-Sarem M, Kissi M. Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University - Computer and Information Sciences 2020; 32(2): 225-231. doi: 10.1016/j.jksuci.2018.05.010

[27] Tutkan M, Ganiz MC, Akyokus S. Helmholtz principle based supervised and unsupervised feature selection methods for text mining. Information Processing & Management 2016; 52 (5): 885-910. doi: 10.1016/j.ipm.2016.03.007

[28] Sarac E, Ozel SA. An ant colony optimization based feature selection for web page classification. The Scientific World Journal 2014; 2014: 649260. doi: 10.1155/2014/649260

[29] Hoque N, Singh M, Bhattacharyya DK. EFS-MI: an ensemble feature selection method for classification. Complex & Intelligent Systems 2018; 4: 105-118. doi: 10.1007/s40747-017-0060-x

[30] Uysal AK. An improved global feature selection scheme for text classification. Expert Systems With Applications 2016; 43: 82-92. doi: 10.1016/j.eswa.2015.08.050

[31] Akın AA, Akın MD. Zemberek, an open source NLP framework for Turkic languages. Structure 2007; 10: 1-5.

[32] Tharwat A. Classification assessment methods. Applied Computing and Informatics 2018; 1: 1-13. doi: 10.1016/j.aci.2018.08.003

[33] Vapnik VN. The Nature of Statistical Learning Theory. Berlin: Springer-Verlag, 1995.

[34] Yadav AK, Chandel SS. Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model. Renewable Energy 2015; 75: 675-693. doi: 10.1016/j.renene.2014.10.046

[35] Pal M. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 2005; 26: 217-222. doi : 10.1080/01431160412331269698