


A New Performance Metric to Evaluate Filter Feature Selection Methods in Text Classification


Rasim Çekik

(Department of Computer Engineering, Faculty of Engineering, Sırnak University, Sırnak, Turkey)

 <https://orcid.org/0000-0002-7820-413X>, rasimcekik@sirnak.edu.tr

Mahmut Kaya*

(Department of Artificial Intelligence and Data Engineering, Faculty of Engineering, Firat University, Elazig, Turkey)

 <https://orcid.org/0000-0002-7846-1769>, mahmutkaya@firat.edu.tr

Abstract: High dimensionality and sparsity are the primary issues in text classification. Using feature selection approaches, the most effective way to solve the problem is to select a subset of features. The most common and effective methods used for this process are filter techniques. Various performance metrics such as Micro-F1, Macro-F1, and Accuracy are used to evaluate the performance of filter methods used for feature selection on datasets. Such methods work depending on a classification algorithm. However, when selecting features in filter techniques, the information on the individual features is evaluated without considering the relationship between the features. In such an approach, the actual performance of the filter technique used in feature selection may not be determined. In such a case, it causes the existing methods to be insufficient in testing the validity of the proposed method. For this purpose, this study suggests a novel performance metric called Selection Error (SE) to determine the actual performance evaluation of filter techniques. The Selection Error metric allows us to analyze the information value of the selected features more accurately than existing methods without relying on a classifier. The feature selection performance of the filtering approaches was performed on six different datasets with both The Selection Error and traditional performance metrics. When the results are examined, it is seen that there is a strong relationship between the proposed performance metric and the classification performance metric results. The Selection Error aims to significantly contribute to the literature by demonstrating the success of filtering feature selection methods, regardless of classifier performance.

Keywords: Text classification, feature selection, filtering methods, performance metric, selection error

Categories: I.2, I.7

DOI: 10.3897/jucs.111675

1 Introduction

Text mining, also known as text analytics, is an artificial intelligence technology that uses natural language processing to make unstructured text in documents available for analysis or machine learning algorithms. It is a review process that extracts new data or answers specific research questions. Text mining, which provides inferences from websites, books, e-mails, articles, online news, and written sources, uses advanced approaches to process and structure data. Text mining tasks include text classification,

text clustering, concept or entity extraction, granular taxonomy modelling, sentiment analysis, entity relationship modelling, and document summarization. The most important of these tasks is text classification. Text classification is the application of classification algorithms to text documents following various pre-processing procedures. The goal is to assign the text document to a group that has previously been categorized or classified based on the document's content. Text classification research covers fields such as document organization, news filtering, spam detection, and opinion mining [Aggarwal and Zhai, 2012]. Processing data in these fields involves additional costs due to the large volume and variety of data. The main reason for these costs is the problem of high dimensionality and sparsity in text classification. One of the most effective ways to find a solution to this problem is to identify a subset of features using a feature selection method. The feature subset is expected to best represent the full feature subset. So, feature selection approaches are a fundamental and effective way to address such challenges and achieve high classification accuracy [Kou et al., 2020].

Each feature used in the classification of text data has a significant impact on classification performance. In large text documents, the number of distinct or varied words can be relatively high [McCallum and Nigam, 1998]. Given this diversity, it is necessary to identify useful features to extract meaningful patterns for classification performance. The goal of feature selection methods is to identify irrelevant features that have little effect on classification or do not contribute to the classification result and to remove these features from the dataset, reducing its size [Forman, 2003]. The selection process helps improve classification performance by reducing noise and irrelevant information in the data.

In the literature, there are three main feature selection approaches: filter, wrapper, and embedded methods [Kaya et al., 2013, Kaya and Bilge, 2016, Çekik and Kaya, 2023]. Filter methods use statistical information to select features without relying on a classifier algorithm. Wrapper methods focus on the relationship between features and use a classification algorithm to identify the most useful feature subspace. Finally, embedded methods, such as decision trees, incorporate feature selection as part of their nature. When evaluating these approaches, statistical methods are effective in terms of data processing and speed.

Filter techniques are advantageous because they use only statistical information to evaluate features, which reduces data processing costs. For this reason, statistical knowledge-based feature selection methods are a suitable solution for high-volume and redundant data. Therefore, the use of filter techniques is becoming more common. The fact that many filter techniques have recently been presented in the literature is an indication of this. In their study, [Çekik and Uysal, 2020], for example, presented proportional rough feature selector (PRFS), a new rough set-based filter technique. Using lower- and upper cluster approaches, the technique provides more precise discrimination of features where there is uncertainty about their distinctiveness. In the [Hancer et al., 2023] paper, they propose an evolutionary filter feature selection approach that can be used for both single and multi-objective scenarios by introducing an objective function inspired by the Neighbourhood Component Analysis (NCA)-based method and then integrating it into the differential evolution framework. [Jin et al., 2023] presented a new filter technique at the level of term and class frequency. In this article, Class Term Frequency (CTF) and Class Document Frequency (CDF) are used to characterize the relevance of terms and categories at levels of term frequency

(TF) and document frequency (DF) levels. Again, [Parlak and Uysal, 2023] presented a new approach called Extensive Feature Selector (EFS). In its calculations, this study employs both corpus-based and class-based probabilities. In this way, examples can be multiplied.

Filtering methods enable the rapid selection of valuable features independent of any classifier algorithm. The selected feature subset consists of individual features with high independent discriminatory power and is expected to provide effective classification performance in text classification. However, since the working strategy of each feature selector is different, a feature may be selected as a distinguishing feature by one feature selector. In contrast, the other may not select this feature. This situation causes each feature selector to offer a different subset of solutions, and accordingly, the problem of determining the best solution from the subsets arises. As a solution to this, using the methods available in the literature, the improvement of the classifier performance for each solution set was evaluated based on criteria such as macro-f1, micro-f1, and accuracy. However, due to the working principles of classifiers, different results can be obtained on the same data set, and the results may vary on a classifier basis. The classifier's performance is also considered the performance of the feature selector. In this case, there needs to be more evaluation of the individual, independent distinctiveness of the features. Furthermore, the idea behind filter selection approaches is to independently select each individual's distinguishing feature. It is not enough to evaluate only the performance of the classifiers to evaluate this idea. This is also the primary reason for the creation of this study. The study proposes a new classifier-independent performance metric called Selection Error (SE) to analyze the impact of filter methods on feature selection. The distance from the $x = y$ line to determine whether the feature belongs to the negative region score is used to calculate SE. The proposed novel performance metric has been tested on many filter techniques and datasets, and very successful results have been obtained.

The rest of the study proceeds as follows: Section 2 contains information on filter feature selection and performance metrics. Section 3 explains the proposed metric. Section 4 describes the experimental work and the results. Section 5 presents the theoretical and managerial implication of the study. Section 6 concludes with a conclusion and discussion.

2 Filter Feature Selection and Performance Metrics

Filter feature selection methods are widely used in machine learning to identify the most relevant features for a given dataset. Filter approaches calculate a score value for each feature based on a statistical calculation. Each score value represents the power of individual discrimination on the classification for each feature.

Information gain (IG) [Yang and Pedersen, 1997] is a widely used and practical feature selection approach in data and text mining. The method is based on Shannon's theory of information and thermodynamics. This approach measures the classification knowledge of a term in any class. In other words, it can be defined as the inverse of entropy. Entropy is the terminology that expresses the disorder of a system. If the number of different values a term can take is high, the IG method that selects this term as a specific term can cause overfitting of the system. Overfitting is a significant disadvantage for IG. The gini index (GI) [Shang et al., 2007] method is an approach

presented to complete the shortcomings of the information gain and gain ratio approaches. In this approach, the entropy value is not used. Instead, GI first calculates each term's class information and gini coefficient. Then, it calculates the gini gain value for each term, depending on the class information relationship of the gini coefficient, and the terms are selected according to this value.

The common and popular Chi-Square (Chi2) approach [Li et al., 2008] is an effective feature selection tool based on statistical information. Chi2 determines whether the relationship between two variables is dependent or independent. Actually, the Chi2 test is a technique used to analyze two independent observations in statistics. Independent observations for text classification are the formation of terms and classes. One of the most effective methods for feature selection in text classification is the distinguishing feature selector (DFS) [Uysal et al., 2012] approach. The DFS determines distinguishing features by removing uninformative aspects and regarding specific term characteristics needs.

A new method for text classification was proposed by Rehman et al. using the balanced accuracy measure, namely the Normalized Difference Measure (NDM) [Rehman et al., 2017]. It is stated that the proposed NDM method does not work efficiently in a skewed dataset consisting of relatively sparse terms. Max-Min Ratio (MMR) [Rehman et al., 2018] gives a high score for each attribute if a term is more frequent in one class but a low score if the term is the same in more than one class. It also completes NDM deficiencies in large and rare data and assigns the highest value to the relevant term.

The deviation from the Poisson distribution (PS) [Ogura et al., 2009] method, derived from the Poisson distribution, is a widely used approach to selecting effective words in information retrieval. The Poisson distribution is a distribution that is commonly used in engineering and statistics, including computer science. PS is a Poisson distribution-based text classification method integrated into feature selection problems.

Other methods include relative discriminant criterion (RDC) [Rehman et al., 2015], variable relative discriminative criterion (MRDC) [Labani et al., 2018], multi-objective relative discriminative criterion (MORDC) [Labani et al., 2020], enhanced gini index (IGI) [Asim et al., 2021], and more.

Filter approaches do not use any classifier or machine learning model for feature selection. However, classification algorithms are used after feature selection to test the success of these methods on datasets. A classifier method is used to evaluate the performance of filter techniques. Text classification is determining which categories a document belongs to according to its content [Dumais et al., 1998, Yang, 1999]. The text classification problem is also known as text categorization. The main purpose of classification is to predict the target class for the specified data accurately. Classifiers such as support vector machines (SVM) [Scholkopf and Smola, 2018], Naïve Bayes (NB) [Rish, 2001], k-nearest neighbors (kNN) [Liao and Vemuri, 2002], and decision tree (DT) [Maimon and Rokach, 2014] are available to perform the classification task. The main purpose of the classifiers, frequently used in the literature, is to maximize classification accuracy. Each process used before classification aims at high classification accuracy.

The effectiveness of filter feature selection methods can be evaluated using various performance metrics, such as accuracy, precision, recall, micro-f1, and macro-f1 [Labani et al., 2018, Labani et al., 2020]. These metrics measure the performance of the

classifier algorithm on the selected feature subset. The advantages and disadvantages of these methods are given in Table 1. Accuracy is the most commonly used performance metric to determine classification performance. The accuracy rate is obtained by dividing the results of the samples that are correct as a result of the classification by the number of all samples.

Another performance metric used to determine classification performance is the precision metric. The precision metric is a metric that shows how many of the positively predicted values are positive. This metric value is significant when the cost of false positive estimates is high. This measured value is obtained by dividing the number of true positive samples (TP) by the sum of the number of true positive samples (TP) and the number of false positive samples (FP). Another performance metric used to determine classification performance is the Recall metric. The Recall metric is a value that shows how many situations that should be predicted positively are predicted. This value is significant when the cost of estimating false negatives is high. This metric value is obtained by the ratio of the number of true positive samples (TP) to the sum of the number of true positive samples (TP) and the number of false negative samples (FN).

Performance Metric	Advantages	Disadvantages
Accuracy	Easily understandable and shows the percentage of correct predictions.	It can be misleading in the case of an unbalanced dataset.
Micro-F1	It suits binary classification problems and provides high precision in multi-class classification problems.	It does not show the difference in performance between different classes, which can be misleading in the case of unbalanced datasets.
Macro-F1	In the case of an unbalanced dataset, it gives more balanced results, showing the difference in performance between different classes.	Because it considers the impact of each class equally, it can be misleading due to the different sizes of the classes.
Precision	It can increase correct predictions by reducing the number of false positives.	It ignores the false negative rate. Therefore, it may not reflect actual performance when not considered together with the false positive rate.
Recall	It can increase correct predictions by reducing false negatives.	It ignores the false positive rate. As a result, it can give misleading results if classifiers fail to predict instances in the rare class correctly.

Table 1: Advantages and disadvantages of classifier performance metrics

Micro-F1 metric, another metric used to determine classification performance, is calculated using the Precision and Recall metric. In calculating this metric, a harmonic mean of the Precision and Recall measures is used to summarize the performance of the classifier algorithm in a more balanced way. One of the best-known F1-measure success measures in the literature is the Macro-F1 metric. The F-measure is calculated for each class in the dataset in the macro average, and all categories are averaged. Therefore, each class is given equal weight regardless of class frequency.

The metrics used to evaluate the performance of filter techniques are directly related to the performance of the classifier method. The classifier's performance can affect the performance of the filter feature selection method, masking the actual performance of the filter method. It would be better to use a classifier-independent

metric to capture the performance of these techniques separately from the classification algorithm. For this reason, a new classifier-independent metric called Selection Error (SE) is proposed in this study.

3 Proposed Metric

The most common method used for feature selection in text classification is filtering. These methods generate a score value indicating the information value of each feature on classification performance. These scores, however, are generated independently of one another, and each feature is evaluated independently of the others. To summarize, filter approaches evaluate each feature independently and separately. These approaches select the N features with the highest distinctiveness and present them to the classifiers by creating a subset of features. The classifier's success on the given feature set is also evaluated as the success of the feature selection approach. However, evaluating all feature spaces together, rather than a single independent feature information, provides a real classifier performance. A high distinctiveness can be obtained when low distinctiveness features are combined with high distinctiveness features. However, filter approaches do not consider such an assessment in their operating principle. Such approaches consider that the selected feature subset consists of independent discriminative features distinctive independent. The feature subset also has high discriminability, which better represents the data. For example, Table 2 shows an example of a simple data scenario taken from Uysal and Gunal's study [Uysal and Gunal, 2012]. The scores of each term according to the DFS feature selection approach are also given in the Table 2.

Document Name	Content	Class	The score of each term according to DFS	
			term	score
D1	iron	A		
D2	iron fan	A		
D3	iron fan	B	iron	0.5000
D4	iron oven	B	fan	0.5714
D5	iron oven mixer	C	oven	0.7000
D6	iron mixer oven	C	mixer	1.0000

Table 2: Sample data and score of each term according to DFS

When Table 2 is examined, it is seen that the term iron occurs equally in all classes, which means that the discriminative power of the term is low. In other words, the term is not distinctive on its own. Similarly, the term fan occurs in two classes and the same number of occurrences. In this case, this term is relatively more discriminative than iron. The term oven appears in both classes B and C, but is more common in class C and thus more distinctiveness than the terms iron and fan. Finally, the term mixer is only mentioned in class C and has a certain majority. Therefore, it is the most distinguishing term. As shown in Table 2, the discrimination of the features is in the order of *mixer* > *oven* > *fan* > *iron* according to DFS. DFS has made the right choice according to the basic working principle of filter approaches. Moreover, the distinctiveness of each term is visibly evident, and DFS has successfully provided this information. However, it is critical to determine how much the classifier demonstrates

this success. Returning to the example, according to the term score ranking of DFS, the triple-term subset $D_f = \{mixer, oven, fan\}$ is expected to provide a more distinguishing classification than the triple subset $D_k = \{mixer, oven, iron\}$. The Multinomial Naive Bayes (MNB) classifier is run on the example for D_f and D_k term subsets and the case is examined. Table 3 shows the mathematical background of the MNB classifier.

$P(C) = \frac{N_c}{N}$	Let $P(C)$ be the class probability, where N is the number of all classes and N_c is the number of class C .
$P(t/C) = \frac{count(t,C)+1}{count(C)+ V }$	$P(t/C)$ is the probability that t terms belong to class C and $ V $ is the number of terms that appear in all documents.
$P(C/D_i) = \prod_{i=0}^m P(t_i/C)$	$P(C/D_i)$, where D_i is the probability that the document belongs to class c and m is the number of terms in the document.

Table 3: The mathematical formula of MNB

Accordingly, the calculation for the subset D_f :

$P(A) = 1/3$	$P(B) = 1/3$	$P(C) = 1/3$
$P(fan/A) = 2/7$	$P(fan/B) = 2/9$	$P(fan/C) = 1/9$
$P(mixer/A) = 1/7$	$P(mixer/B) = 1/9$	$P(mixer/C) = 3/9$
$P(oven/A) = 1/7$	$P(oven/B) = 3/9$	$P(oven/C) = 2/9$
$P(A/D_f) = 0.0019$	$P(B/D_f) = 0.0027$	$P(C/D_f) = 0.0027$

The calculation for the subset D_k :

$P(A) = 1/3$	$P(B) = 1/3$	$P(C) = 1/3$
$P(iron/A) = 3/7$	$P(iron/B) = 3/9$	$P(iron/C) = 3/9$
$P(mixer/A) = 1/7$	$P(mixer/B) = 1/9$	$P(mixer/C) = 3/9$
$P(oven/A) = 1/7$	$P(oven/B) = 3/9$	$P(oven/C) = 2/9$
$P(A/D_k) = 0.0029$	$P(B/D_k) = 0.0041$	$P(C/D_k) = 0.0082$

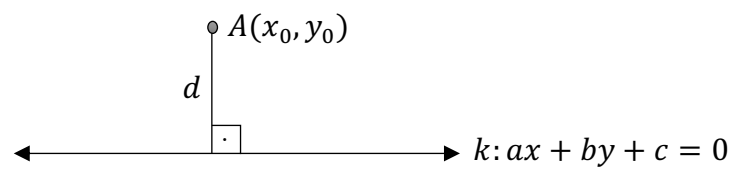
When the results are evaluated, it is seen that a more distinguishing classification process is realized with D_k . For the MNB classifier, this means that the terms in D_k are classified more discriminatively than the terms in D_f . However, DFS indicates that there will be a more distinguishing classification with D_f . It demonstrates that when terms with low discriminative power alone are considered alongside other terms, classifiers perform better. As a result, assessing the performance of filter feature approaches only based on classifier performance is insufficient. An evaluation metric appropriate for the study purposes is required to see the actual performance of filter approaches. This article fills that void by proposing a new evaluation metric known as Selection Error (SE).

In this study, the proposed performance metric groups low discriminative features rather than high discriminative features and calculates how many features are selected

from this group in feature selection. The distribution of a feature across classes provides important information about its distinctiveness. The $x = y$ line is used to extract this information. For example, let $N(x_0, y_0)$ be a point for a dataset of binary classes C_1 and C_2 . x_0 denotes the frequency of occurrence of the term in class C_1 and y_0 in class C_2 . Given points $A(100, 20)$ and $B(100, 100)$ with distinctiveness $B < A$, point A has a passing frequency of 100 in class C_1 and 20 in class C_2 . However, point B passes equally in both classes. Point B is right on the $x = y$ line, while point A is far away. As a result, the $x = y$ line can be used to draw a boundary for distinguishing features, and it was used in this study.

In the two-dimensional plane where the feature belongs to one class (positive) and the feature belongs to the other class (negative) on the axis, regions close to the $x = y$ line mean the feature has low distinctiveness. In other words, if a feature frequently occurs in all classes, it is always close to the $x = y$ line. Therefore, its distinctiveness is low. Also, if a feature rarely occurs in one class and not in the other classes, its discrimination is low. However, some features have relatively low discrimination. However, some features located close to the $x=y$ line have relatively low discrimination. In this study, this area is defined as the negative region. The boundary of the negative region is defined by a threshold value such as α . The region below α is defined as the negative region. The α value is a threshold value, a boundary ratio that determines the area of the negative region. This value is also an indicator of the distribution of a feature across classes. The value is randomly selected in the range $[0, 0.2]$, and a value in this range indicates that the feature is 80%-100% distributed across classes. This means that the frequency of occurrence of the term in a single class is at most 20%. This rate means that the distinctiveness of a feature is low. Therefore, the discrimination quality of a feature is expected to be above this threshold. The opposite situation means that the feature has low discriminative quality.

In this study, selecting a feature from the negative region of a feature selection approach is defined as the Selection Error (SE). To determine the Se, it is necessary to first calculate the features belonging to the negative region score (**Feature Quality – FQ**). For this, first, the distance to the $x = y$ line is calculated. The formula for the distance of a point to a line is used for this calculation. Accordingly, the distance of $t(x_0, y_0)$ to the line: $ax + by + c = 0$;



$$d = \frac{|a * x_0 + b * y_0 + c|}{\sqrt{a^2 + b^2}} \tag{1}$$

obtained by the equation 1. Since the line used is $x = y$, the distance of a point to this line is as follows:

$$d = \frac{|\mathbf{x}_0 - \mathbf{y}_0|}{\sqrt{2}} \tag{2}$$

In equation 2, \mathbf{x}_0 indicates belonging to the class and \mathbf{y}_0 indicates non-belonging. To indicate these situations in the article, \mathbf{c}_i will be used instead of \mathbf{x}_0 for belonging to the class, and \mathbf{c}_j will be used instead of \mathbf{y}_0 for non-belonging. The \mathbf{d} information alone is insufficient to obtain the information that the quality of the features is low. Therefore, some additional information is needed. Knowing the distribution of features according to classes by looking at the density of a feature in the whole feature set also provides important information on discrimination. For this operation, the logarithm of the distance \mathbf{d} is taken based on the number of features in the dataset (\mathbf{fs}). In addition, the number of documents in the dataset is another piece of information to be used to obtain a more effective calculation. The frequency of a feature in documents and its frequency by class allows us to distinguish between features. Based on this information, the **feature quality (FQ)** of a feature can be calculated within the equation 3:

$$FQ(t) = \sum_{i=0}^{m-1} \sum_{j=i+1}^{m-1} \left(\frac{d_{\mathbf{c}_i \mathbf{c}_j}}{\mathit{count}(\mathbf{c}_i + \mathbf{c}_j) + 1} \right) * \log_{\mathbf{fs}}^{|\mathit{pr}_{\mathbf{c}_i} - \mathit{fr}_{\mathbf{c}_j}|} \tag{3}$$

and, $\log_{\mathbf{fs}}^{|\mathit{pr}_{\mathbf{c}_i} - \mathit{fr}_{\mathbf{c}_j}|} = \begin{cases} 0.01 & \text{if } |\mathit{pr}_{\mathbf{c}_i} - \mathit{fr}_{\mathbf{c}_j}| = 0 \text{ or } 1 \\ \log_{\mathbf{fs}}^{|\mathit{pr}_{\mathbf{c}_i} - \mathit{fr}_{\mathbf{c}_j}|} & \text{otherwise.} \end{cases}$

In the equation 3,

FQ(t): Score for belonging to the negative region for feature \mathbf{t} .

m: Number of classes.

$d_{\mathbf{c}_i \mathbf{c}_j}$: Distance of the feature to the $\mathbf{x} = \mathbf{y}$ line at class coordinates \mathbf{c}_i and \mathbf{c}_j .

count($\mathbf{c}_i + \mathbf{c}_j$): Sum of the frequency of occurrence of the feature in classes \mathbf{c}_i and \mathbf{c}_j .

fs: All feature size

$\mathit{pr}_{\mathbf{c}_i}$: Frequency of occurrence of the feature in class \mathbf{c}_i

$\mathit{fr}_{\mathbf{c}_j}$: Frequency of occurrence of the feature in class \mathbf{c}_j

At this point, it is helpful to underline the following. **FQ** may not be sufficient to distinguish high distinctiveness features in regions far from the $x = y$ line. Moreover, **FQ** does not have any idea in this direction. Identifying and sorting highly distinctive features is the task of feature selection approaches. The **FQ** grouped only the low distinctiveness features, and a lower bound was set between the good and bad distinctiveness of the features.

First, the proportion of each term belonging to the FQ negative region is calculated to calculate the SE values for each term in the text feature set given in Table 2. According to this;

iron

$d_{c_i c_j}$	$\left(\frac{d_{c_i c_j}}{\text{count}(c_i + c_j) + 1}\right) * \log_{f_s}^{ \text{pr}_{ci} - \text{fr}_{ci} }$
$d_{AB} = \frac{ 2 - 2 }{\sqrt{2}} = 0$	$\frac{0}{(2 + 2) + 1} * \log_4^{ 2-2 } = 0 * 0.01 = 0.0000$
$d_{AC} = \frac{ 2 - 2 }{\sqrt{2}} = 0$	$\frac{0}{(2 + 2) + 1} * \log_4^{ 2-2 } = 0 * 0.01 = 0.0000$
$d_{BC} = \frac{ 2 - 2 }{\sqrt{2}} = 0$	$\frac{0}{(2 + 2) + 1} * \log_4^{ 2-2 } = 0 * 0.01 = 0.0000$
$FQ(\text{iron}) = 0 + 0 + 0 = 0.0000$	

fan

$d_{c_i c_j}$	$\left(\frac{d_{c_i c_j}}{\text{count}(c_i + c_j) + 1}\right) * \log_{f_s}^{ \text{pr}_{ci} - \text{fr}_{ci} }$
$d_{AB} = \frac{ 1 - 1 }{\sqrt{2}} = 0$	$\frac{0}{(1 + 1) + 1} * \log_4^{ 1-1 } = 0 * 0.01 = 0$
$d_{AC} = \frac{ 1 - 0 }{\sqrt{2}} = \frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}/2}{(1 + 0) + 1} * \log_4^{ 1-0 } = \frac{\sqrt{2}}{4} * 0.01 = 0.0035$
$d_{BC} = \frac{ 1 - 0 }{\sqrt{2}} = \frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}/2}{(1 + 0) + 1} * \log_4^{ 1-0 } = \frac{\sqrt{2}}{4} * 0.01 = 0.0035$
$FQ(\text{fan}) = 0 + 0.0024 + 0.0024 = 0.0070$	

oven

$d_{c_i c_j}$	$\left(\frac{d_{c_i c_j}}{\text{count}(c_i + c_j) + 1}\right) * \log_{f_s}^{ pr_{ci} - fr_{ci} }$
$d_{AB} = \frac{ 0 - 1 }{\sqrt{2}} = \frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}/2}{(1 + 0) + 1} * \log_4^{ 0-1 } = \frac{\sqrt{2}}{4} * 0.01 = 0.0036$
$d_{AC} = \frac{ 0 - 2 }{\sqrt{2}} = \sqrt{2}$	$\frac{\sqrt{2}}{(0 + 2) + 1} * \log_4^{ 0-2 } = \frac{\sqrt{2}}{3} * 0.5 = 0.2357$
$d_{BC} = \frac{ 1 - 2 }{\sqrt{2}} = \frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}/2}{(1 + 2) + 1} * \log_4^{ 1-2 } = \frac{\sqrt{2}}{8} * 0.01 = 0.0018$

$$FQ(\text{oven}) = 0.0036 + 0.2357 + 0.0035 = 0.2410$$

mixer

$d_{c_i c_j}$	$\left(\frac{d_{c_i c_j}}{\text{count}(c_i + c_j) + 1}\right) * \log_{f_s}^{ pr_{ci} - fr_{ci} }$
$d_{AB} = \frac{ 0 - 0 }{\sqrt{2}} = 0$	$\frac{0}{(0 + 0) + 1} * \log_4^{ 0-2 } = 0 * 0.5 = 0.0000$
$d_{AC} = \frac{ 0 - 2 }{\sqrt{2}} = \sqrt{2}$	$\frac{\sqrt{2}}{(0 + 2) + 1} * \log_4^{ 0-2 } = \frac{\sqrt{2}}{3} * 0.5 = 0.2357$
$d_{BC} = \frac{ 0 - 2 }{\sqrt{2}} = \sqrt{2}$	$\frac{\sqrt{2}}{(0 + 2) + 1} * \log_4^{ 0-2 } = \frac{\sqrt{2}}{3} * 0.5 = 0.2357$

$$FQ(\text{mixer}) = 0 + 0.2357 + 0.2357 = 0.4714$$

After calculating the **FQ**, the **SE** of each feature is determined finally. SE is calculated within the equation 4:

$$SE = \frac{\text{size}(FQ < \alpha)}{\text{size}(FQ)}, \quad \alpha \in [0, 0.2] \tag{4}$$

A feature in the negative region has low discrimination. Therefore, the main task of filter approaches is to select as few features as possible from this region. If the feature selection approach chooses less of the desired number of features in the negative region, it can be said that its performance is better. Therefore, this study presents a metric to

help evaluate the performance of filter feature selection approaches without any classifier.

Returning to the example, **SE** for the DFS selector when $\alpha = 0.2$. Order of selecting DFS features if remembered: **mixer > oven > fan > iron** was as follows. Accordingly;

$$\text{If } N = 2; \text{selected feature} = \{\text{mixer} > \text{oven}\} \text{ SE} = \frac{\text{size}(FQ < 0.2)}{\text{size}(FQ)} = \frac{0}{2} = 0$$

$$\text{If } N = 3; \text{selected feature} = \{\text{mixer} > \text{oven} > \text{fan}\} \text{ SE} = \frac{\text{size}(FQ < 0.2)}{\text{size}(FQ)} = \frac{1}{3} = 0.3333$$

$$\begin{aligned} \text{If } N = 4; \text{selected feature} &= \{\text{mixer} > \text{oven} > \text{fan} > \text{iron}\} \text{ SE} \\ &= \frac{\text{size}(FQ < 0.2)}{\text{size}(FQ)} = \frac{2}{4} = 0.5000 \end{aligned}$$

Figure 1 shows a 3D graphical representation of the properties and the negative region of the given sample for $\alpha = 0.0070$.

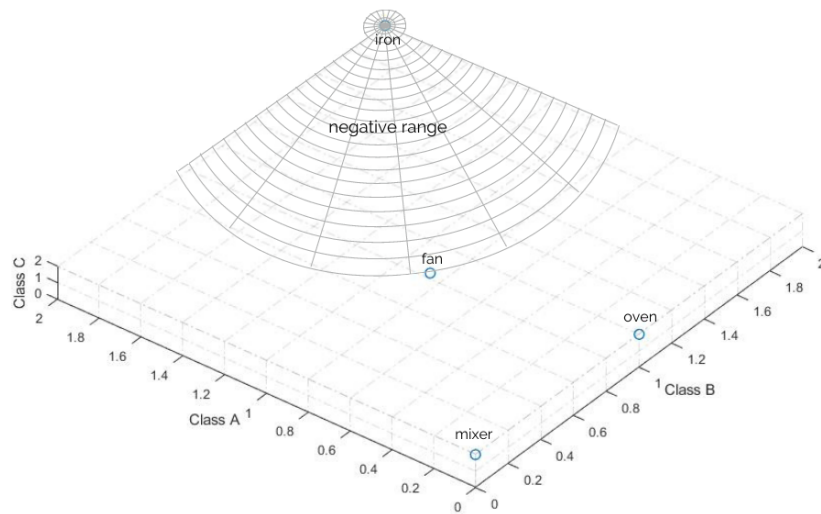


Figure 1: Sample negative region for the features in Table 2

The most crucial difference between the existing performance metrics and SE: Traditional performance metrics require a classifier to evaluate the selected feature subset in feature selection approaches, and the performance of any classifier is also the success of the proposed feature selection approach. However, SE evaluates the feature subset chosen without presenting it to any classifier. This is important to show the

consistency of a feature selection approach on different datasets. At the same time, it provides the opportunity to evaluate the approach together with other metrics more healthily and efficiently. The score values in Figure 2 are representative. The values are randomized.

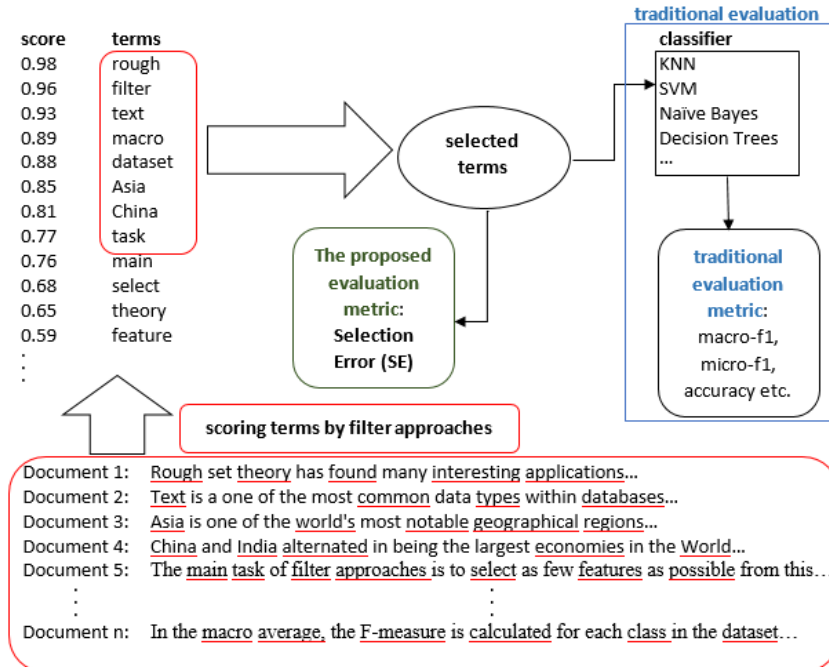


Figure 2: The difference between traditional evaluation criteria and SE in filter feature selection approaches

As a result, a low value of the proposed metric for a feature selection method means that the selection method has a high ability to select features with few errors. Conversely, it means that the selection technique does not fulfil the purpose of the selection strategy. In other words, a high value of the proposed metric means that the filter feature selection method does not select features independently and individually from the most discriminative features.

4 Experimental Results

In this section, the results of the experimental studies will be presented. First, brief information about the datasets used in the article is given. Then, the effect of the SE performance metric on the success of feature selection approaches is analyzed. For this purpose, the performances of the feature selection approaches described in the previous sections are compared. In addition, the performance of the classifiers using the features selected by the feature selection algorithms is also shown in this section. Finally, some

statistical analyzes are included in this section as a table to indicate whether the SE is statistically significant. Another important point underlined in this section is that Macro-F1 and Micro-F1 show more meaningful results, as the Precision and Recall metric also affect Macro-F1 and Micro-F1. Therefore, it can be said that the most important metrics among the approaches compared with SE are Macro-F1, Micro-F1 and Accuracy metrics. These metrics are mainly taken into account in experimental comparisons.

4.1. Datasets

This study used six different datasets to evaluate the performance of feature selection methods. The first used dataset is SMS data [Nuruzzaman et al., 2011, Uysal et al., 2013], consisting of 425 spam and 450 non-spam Short Message Service (SMS) text messages. Another dataset, the Youtube Spam Collection, combines five similar datasets (Psy, KatyPerry, LMFAO, Eminem, Shakira) containing spam and non-spam reviews of various Youtube videos. The dataset has five datasets of 1,956 actual messages. PSY, KatyPerry, LMFAO, Eminem, and Shakira video comntaries [Alberto et al., 2015]. The dataset created in the study was named All Youtube Spam Collection (AYSC). Another dataset used in this study is a subset of the Sentiment140 dataset called Sentiments [Go et al, 2009]. The fourth dataset is the Amazon Reviews for Sentiment Analysis (ARfSA) [Amazon Reviews, 2020]. The dataset used in this study is a subset of ARfSA, which consists of text documents containing the emotions of millions of Amazon customers. Enron, Email Dataset is the other dataset used in this study [Enron Email, 2015]. The dataset includes email data from approximately 150 users, mostly senior management of the Enron organization. The dataset was collected and prepared by the CALO Project (A Learning and Organizing Cognitive Assistant) and contains approximately 0.5 million messages in total. However, a subset of it, Enron1, was used in this study. Recently used dataset Reuters-21578 dataset is a collection of documents containing news articles. The original word has 10,369 documents and a vocabulary of 29,930 words. The dataset is a collection of documents that appeared on the Reuters news portal in 1987. Documents were aggregated and indexed by category [Reuters-21578, 1997].

4.2. Analysis of Selection Error

The SE metric aims to present an analysis study by detecting the selected features with very low discrimination. With the help of this analysis, it can be easier to see the errors of selection approaches in the selection method. It can also guide evaluating the working principle of selection approaches. For example, the best ten features identified by feature selection on six datasets are given in Table 4. However, some of these selected features are not meaningful enough to represent the dataset. The features that should not be chosen with weak and bad discrimination are shown in bold in the Table 4.

No.	Terms							
	1	2	3	4	5	6	7	8
(a)								
DFS	call	free	txt	stop	repli	mobil	www	min
CHI2	call	free	txt	stop	repli	text	min	send
IG	call	free	txt	stop	repli	min	send	text
GI	call	free	txt	stop	text	repli	mobil	www
PS	call	free	txt	repli	stop	text	min	send
MMR	mobil	msg	nokia	offer	custom	land	rcvd	loyalti
NDM	mobil	msg	nokia	offer	custom	land	rcvd	loyalti
(b)								
DFS	sad	followfridai	miss	love	sigh	hate	suck	quot
CHI2	sad	miss	followfridai	love	sigh	hate	suck	quot
IG	sad	followfridai	miss	love	sigh	hate	suck	musicmondai
GI	quot	love	good	miss	http	work	sad	followfridai
PS	quot	sad	followfridai	miss	love	sigh	hate	suck
MMR	rat	apl	sad	present	mean	sundai	guess	movi
NDM	rat	apl	sad	present	mean	sundai	quot	guess
(c)								
DFS	great	wast	disappoint	monei	love	bad	excel	worst
CHI2	great	wast	disappoint	love	monei	bad	excel	worst
IG	great	wast	disappoint	love	monei	bad	worst	excel
GI	great	book	good	love	time	don	read	work
PS	book	great	movi	wast	disappoint	love	monei	bad
MMR	fals	huh	trashi	fad	millionair	garbag	hotboi	cash
NDM	fals	huh	trashi	fad	millionair	garbag	hotboi	cash
(d)								
DFS	enron	cc	hpl	gas	Ect	daren	hou	pm
CHI2	http	cc	enron	gas	Ect	pm	meter	forward
IG	Cc	gas	ect	pm	Meter	http	corp	volum
GI	sbjct	enron	cc	hpl	Gas	forward	ect	daren
PS	ect	hou	enron	meter	Deal	subject	gas	pm
MMR	medic	sleep	fda	physician	Ill	restor	bone	emot
NDM	sleep	fda	physician	ill	restor	bone	emot	disear
(e)								
DFS	check	youtub	subscrib	video	channel	song	monei	gui
CHI2	check	youtub	video	subscrib	song	channel	gui	hey
IG	youtub	check	video	song	subscrib	gui	hey	work
GI	check	youtub	video	subscrib	song	channel	love	monei
PS	check	youtub	video	br	subscrib	song	channe	gui
MMR	googl	type	open	smoke	guruofmovi	uncl	like	check
NDM	googl	type	open	smoke	guruofmovi	uncl	check	video
(f)								
DFS	cts	wheat	net	oil	shr	tonn	corn	barrel
CHI2	cts	net	shr	qtr	rev	loss	acquir	profit
IG	cts	net	wheat	bank	shr	qtr	tonn	export
GI	cts	net	shr	wheat	oil	barrel	qtr	march
PS	mln	dlr	cts	loss	net	bank	pct	billion
MMR	sorghum	wheat	grain	oat	bread	veget	cwt	bu
NDM	sorghum	wheat	oat	bread	veget	cwt	grain	bu

Table 4: Top-8 features in SMS (a), Sentiments (b), Amazon_RfSA (c), Enron1 (d), AYSC (e), Reuters-21578 (f) using FS methods

Table 5 shows the number of mentions of weak and non-weak features in the document in Table 4 for some datasets. In this table, the features with poor discrimination are more readily visible, and it is easier to understand why they have poor discrimination.

Distinctive terms					Not distinctive terms				
(a)									
	sigh	miss	love	sad	apl	present	mean	sundai	movi
C1	305	669	369	471	1	12	53	42	97
C2	29	154	1031	18	0	20	33	58	121
	work	good	hate		don				
C1	701	428	292		453				
C2	379	866	46		494				
(b)									
	great	wast	love	bad	trashi	fad	sever	work	movi
C1	1203	1122	1401	1240	7	6	8	1761	1351
C2	3358	35	295	292	3	3	16	1553	987
	monei	poor	don		time	millionair	hotboi	book	
C1	1322	595	1972		2175	7	2	2741	
C2	193	54	911		1897	6	0	2900	
(c)									
	enron	cc	hpl	ect	forti	thick			
C1	0	6	0	8	4	2			
C2	1015	891	761	766	1	1			
	meter	sleep	forward						
C1	1	18	53						
C2	591	1	803						
(d)									
	check	youtub	video	subscrib	uncl	guruofmovi	br		
C1	392	216	234	182	1	1	50		
C2	0	3	33	0	0	1	37		
	googl	type	hey						
C1	22	17	57						
C2	0	0	2						

Table 5: Sentiments (a), Amazon_RfSA (b), Enron1 (c), AllYoutube (d)

When Table 5 is examined, the term "apl" chosen for the sentiments dataset was not used once in C1 and never in C2. Using a term in only one class is, of course, important in terms of distinctiveness. But it is even more important that it be at a certain frequency. Therefore, if a term is used in a single class and has a certain number (frequency) in that class, its distinctiveness is high. However, the passing frequency of "apl" is extremely low. Therefore, the number of these examples given can be increased even more. As another example, the term "work" for the Amazon_RfSA dataset occurs almost in close numbers in both classes. Therefore, its discrimination is also low. To see the situation more clearly, the representation of the terms given in Table 5 in the coordinate plane for Sentiments and Amazon_RfSA data sets are presented in Figure 3 and Figure 4, respectively.



Figure 3: Distributions of terms for Sentiments dataset

The terms in the regions close to the $x = y$ line in Figure 3 are terms with low discrimination. Moreover, the terms on the line have the lowest discrimination (zero). Similarly, the terms with low discrimination for the Amazon_RfSA dataset are given in Figure 4. These terms also have high SE values. Therefore, these terms also have high SE values. Consequently, selecting these terms should be outside the scope of feature selection approaches. If any feature selection approach can choose between these terms, the approach's selection method is said to be misleading.

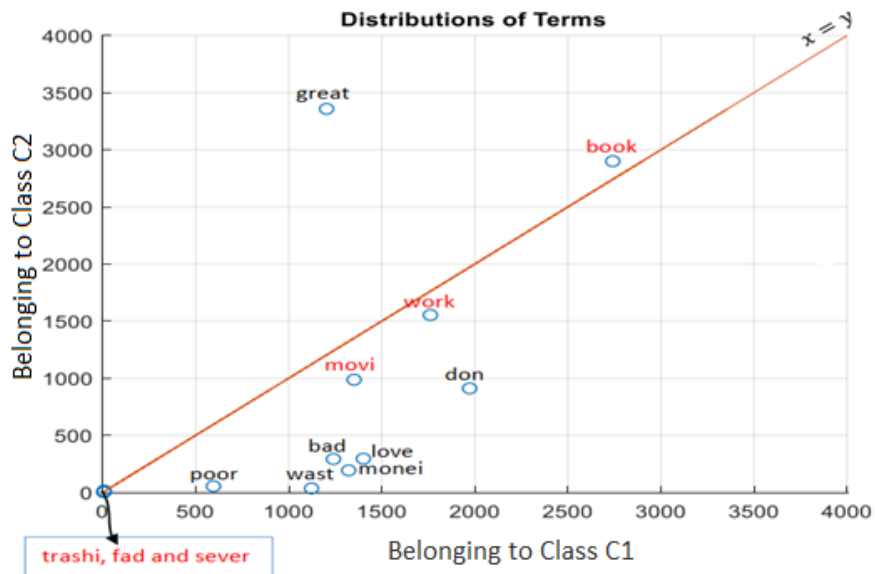


Figure 4: Distributions of terms for the Amazon_RfSA dataset

The process of diagnosing epileptic seizures by using EEG data is obtained as a result of combining various steps. If this process is to be done through a machine learning classifier, while making it ready for the classifier with the help of various pre-processes and then feature extraction methods; If this diagnostic process is performed with deep learning methods, the feature extraction process changes according to the path followed by the researcher who carried out the study. The transfer learning process, which deep learning algorithms have performed due to its layered structure, already performs the feature extraction process within its own structure. However, by using various transformation algorithms, the available data can be used to increase the learning success of deep learning methods.

4.3. Results

As previously stated, numerous performance metrics evaluate the performance of feature selection approaches. The macro-f1 metric, on the other hand, is the most commonly used performance metric in text mining. This study provides a performance comparison of filtering techniques on this metric.

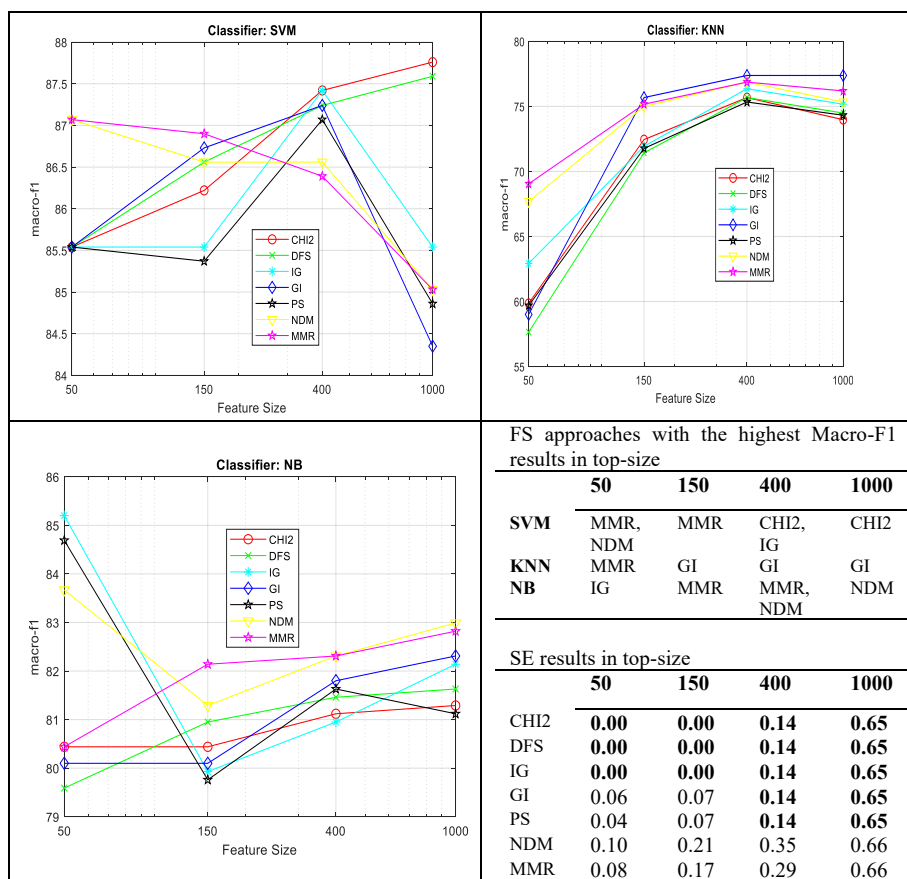


Figure 5: Macro-F1 results for AYSC Dataset using SVM, KNN, NB

Figure 5-10 depicts the performance results of the feature selection methods. An additional table contains the name and SE value of the feature selection methods that exhibit the best performance in all dimensions in Figures. When Figure 5 is examined, DFS, CHI2, and IG gave the best results for the SE metric in the 50 and 150 dimensions, while the DFS, CHI2, IG, GI, and PS approaches succeeded in the 400 and 1000 dimensions. In this case, it can be said that DFS, CHI2, and IG successful and balanced results have been achieved in all dimensions of the ARfSA dataset. However, as seen in Figure 5, it is difficult to give a specific opinion on filter feature selection methods according to the Macro-F1 metric. For example, when the SVM classifier is used in dimension 50, NDM and MMR approaches give the best results for all metrics except SE. At the same time, this situation changed when the NB classifier was used, and IG gave the best results for the Macro-F1 metric. In addition, the results were analyzed for Micro-F1 and Accuracy metrics. IG provides the best results for Micro-F1 and Accuracy metrics, IG and PS for Precision metric, and MMR for Recall metric gave the best results. As seen in the analysis, the information shows that the performance of feature selection approaches changes as the classifier changes.

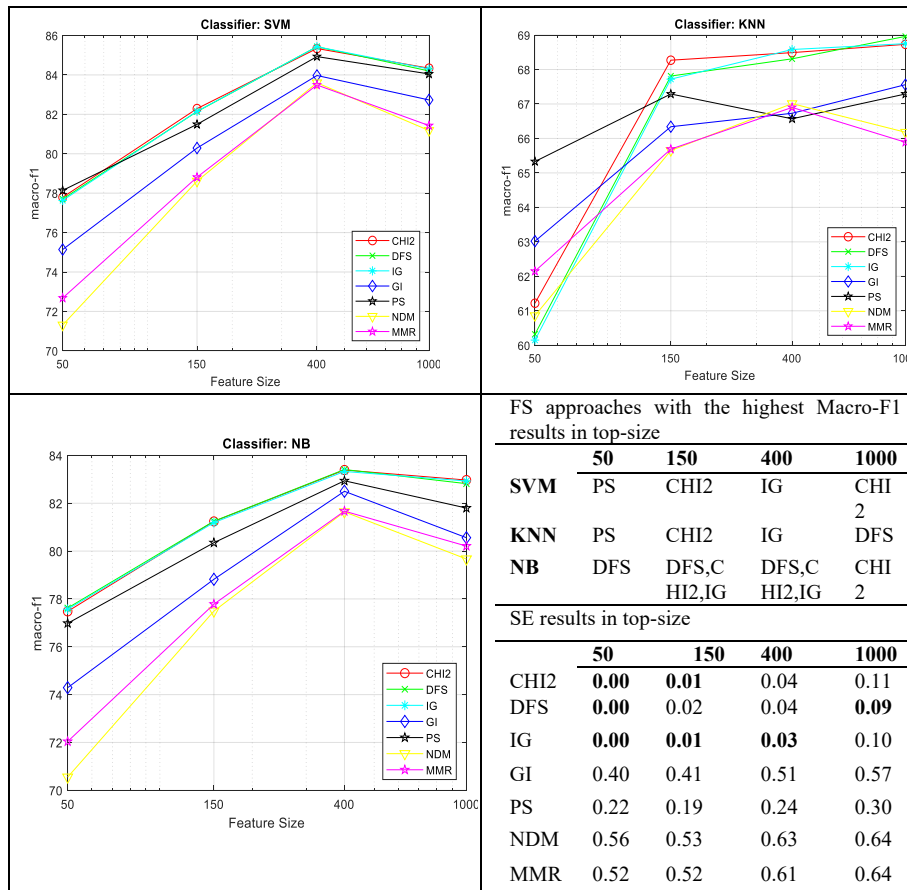


Figure 6: Macro-F1 results for ARfSA Dataset using SVM, KNN, NB

When Figure 6 is analyzed regarding SE, the best results in dimension 50 were achieved with the CHI2, DFS, and IG approaches. While the best success was achieved with IG in dimension 400, the best result was achieved with DFS in dimension 1000. From this information, it can be concluded that for the ARfSA dataset, the IG approach chooses better features overall. Looking at the other metric, PS performed best in dimension 50 when the SVM classifier was used, while CHI2 showed the best performance in dimensions 150 and 1000. Also, IG achieved the best results in dimension 400. It is seen that this situation changes depending on the classifier algorithm. For example, the DFS rather than the CHI2 approach showed the best performance for dimension 50 when using the NB classifier. For dimension 1000, the CHI2 approach gave better performance than the DFS approach.

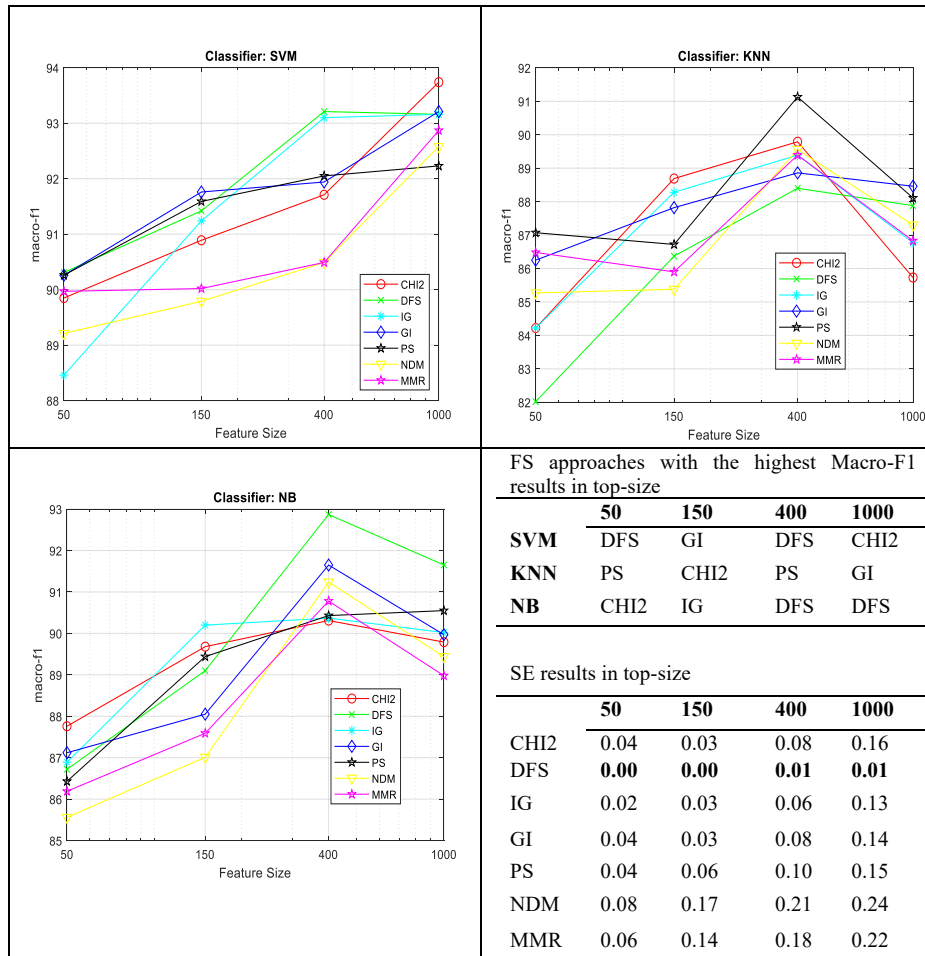


Figure 7: Macro-F1 results for Enron-1 Dataset using SVM, KNN, NB

When Figure 7 is analyzed in terms of SE metric, it is seen that the DFS approach achieves the best results in all dimensions. For other performance metrics, it is also possible to see that DFS performs well. This information also shows a strong correlation between SE and the selection performance of filtering approaches.

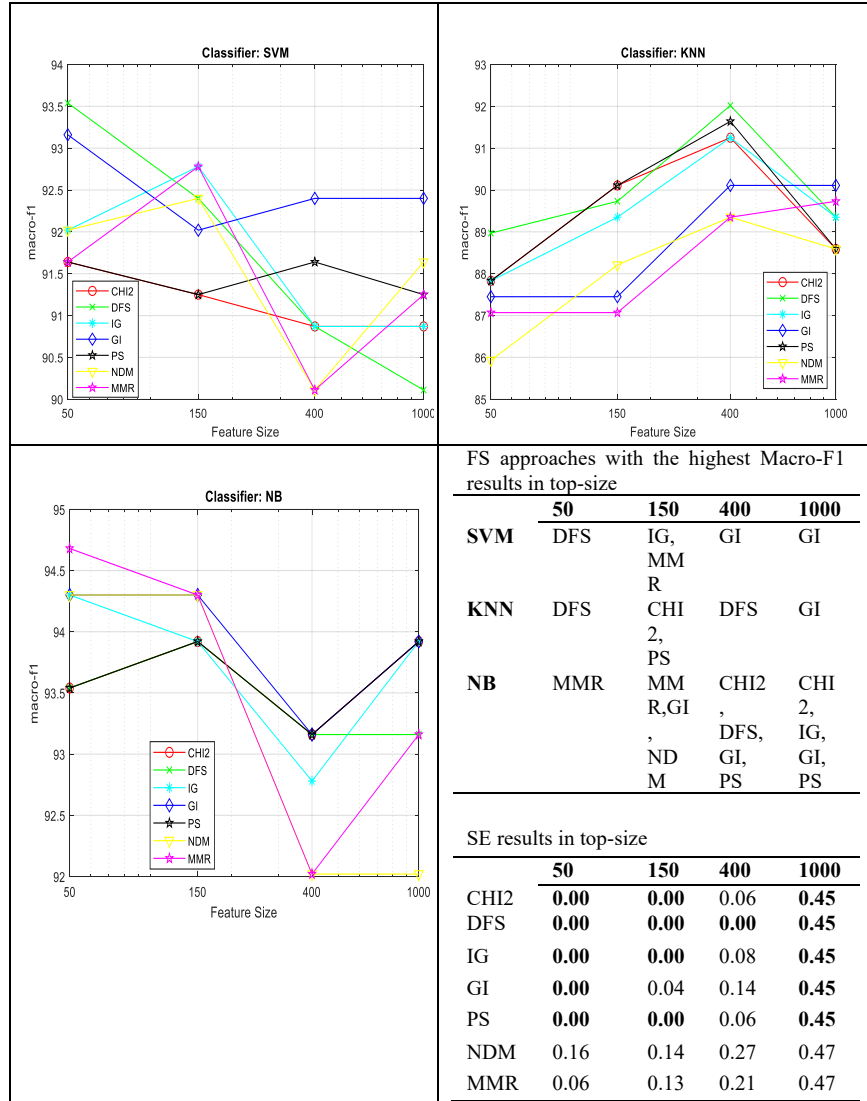


Figure 8: Macro-F1 results for SMS Dataset using SVM, KNN, NB

When Figure 8 is analyzed in terms of SE metric, it is seen that the DFS approach achieves the best results in all dimensions. According to this information, DFS has shown an effective feature selection success on the SMS dataset. In addition, we see that CHI2, IG, GI, and PS approaches to achieve good performance in the 1000th

dimension. In this sense, when we analyze the results, we see that the proposed metric achieves effective results in performance evaluation.

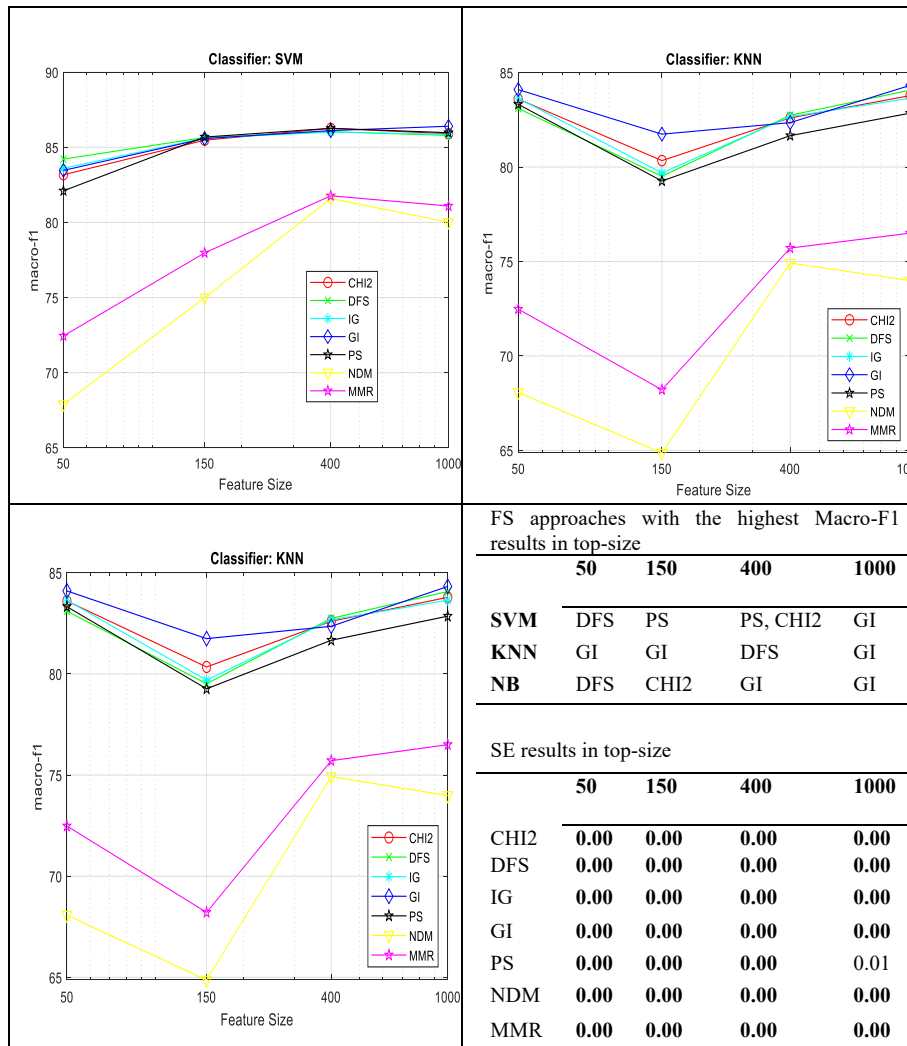


Figure 9: Macro-F1 results for Reuters_21578 Dataset using SVM, KNN, NB

Analyzing Figure 9 shows that all filter selection methods achieve outstanding results regarding the SE metric on the Reuters_21578 datasets. When we examine Figure 10, we see that CHI2, DFS, IG, and PS methods have been selected well according to the SE metric in dimension 50 on the Sentiments dataset. Also, in the 400th and 1000th dimensions, the DFS method shows the best selection according to the SE metric. These results are similar to the classification performance results.

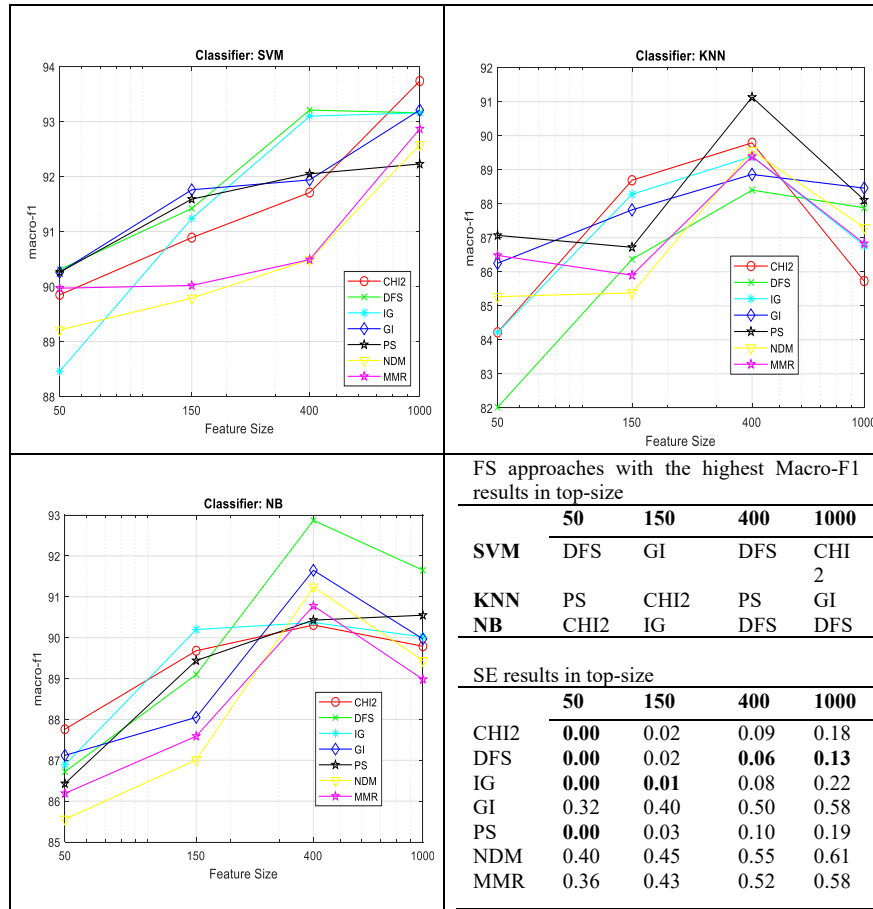


Figure 10: Macro-F1 results for Sentiments Dataset using SVM, KNN, NB

The proposed novel performance metric has achieved perfect results on each data set in measuring the performance of filtering methods. The average results of the classification methods on different datasets also provide valuable information to more clearly see the true strength of the proposed novel performance metric and obtain consistent information. Any filtering feature selection approach that works well on any dataset may not yield good results on another dataset. For this purpose, Table 6 was created to see the average results of the feature selection methods on all data sets, and the average results of each filtering method are given based on dimensions. Table 6 shows that the SE metric proposed to evaluate the performance of the filtering methods is generally quite successful. In general, filtering methods with low SE performance metrics also obtained high Macro-F1 results in classification results. When SE and Macro-F1 performance metric results are evaluated together, it is seen that the DFS method selects the features with less error than other methods. As a result, the DFS approach has a more stable and balanced feature selection mechanism than other filtering methods. As can be seen from the results, the proposed novel performance

metric will significantly contribute to the literature to see the effects of statistical feature selection methods independent of classifier performance.

FS	Feature Size=50				Feature Size=150			
	SVM	KNN	NB	SE	SVM	KNN	NB	SE
CHI2	82.42	71.18	81.50	0.01	84.17	76.10	82.97	0.01
DFS	82.94	70.28	82.24	0.00	84.51	75.32	83.11	0.01
IG	82.29	71.36	81.12	0.00	84.36	75.75	82.82	0.01
GI	82.13	71.30	79.35	0.14	84.05	75.70	82.14	0.16
PS	82.37	72.13	81.18	0.05	84.04	75.44	82.58	0.06
NDM	77.46	69.08	75.84	0.22	81.60	72.55	79.12	0.25
MMR	78.54	70.49	75.82	0.18	82.30	73.21	79.60	0.23
FS	Feature Size=400				Feature Size=1000			
	SVM	KNN	NB	SE	SVM	KNN	NB	SE
CHI2	85.19	75.99	83.29	0.07	85.73	77.20	82.85	0.26
DFS	85.21	76.63	83.88	0.04	85.60	77.08	83.63	0.22
IG	85.40	76.49	83.61	0.07	85.27	77.26	82.91	0.26
GI	85.23	77.06	83.39	0.23	85.05	76.66	83.41	0.40
PS	85.21	76.08	83.27	0.11	85.07	76.08	82.98	0.29
NDM	83.18	74.10	80.91	0.33	83.71	74.10	81.74	0.44
MMR	83.30	74.69	81.54	0.30	83.77	74.69	81.84	0.43

Table 6: Average Macro-F1 and SE Results of ALL Dataset using SVM, KNN, NB

5 The Theoretical and Managerial Implication of the Study

Feature selection methods are frequently used in the literature to solve classification problems in machine learning. These approaches are preferred to identify the features that have the ability to better represent any dataset. The selection method is aimed to increase the classification success on the dataset and reduce the running time of the algorithm. For this purpose, filter feature selection methods used in text classification problems with high data size provide successful solutions. The success of filter feature selection methods has a direct impact on classification performance. In this study, an SE metric is proposed to evaluate the performance of filter feature selection before the classification process. The theoretical and managerial implications of the proposed method are discussed in the following subsections.

5.1. Theoretical Implications

In the literature, a classifier algorithm is used to evaluate the performance of filter algorithms. Micro-F1, Macro-F1, and Accuracy are performance metrics that combine the performance of the filter methods with the performance of the classifier method. The SE metric introduced in this research extends beyond previous metrics for evaluating feature selection methods and provides a more comprehensive knowledge of these methods' effectiveness independent of the classifier.

Because the Selection Error measure is not dependent on a specific classification technique, it eliminates the bias associated with classifier-based choices. This contributes to the theoretical foundation by stressing the independence of feature selection evaluation from the complexities of a specific classification method, making the suggested measure more adaptable and broadly applicable.

5.2. Managerial Implications

The study provides a valuable metric, Selection Error, for practitioners and decision-makers involved in text classification tasks. This metric enables a more informed selection of features by offering a comprehensive evaluation that goes beyond conventional metrics. This gives researchers the ability to make strategic decisions about feature selection approaches that are adapted to their individual settings.

The strong correlation observed between the Selection Error and traditional classification performance metrics indicates the robustness and generalizability of the proposed metric. This information can be used by researchers to improve the generalization capabilities of their models, ensuring that the selected characteristics contribute meaningfully to overall classification performance across varied datasets. This increases the efficiency and efficacy of text classification applications.

6 Conclusion and Discussion

In this study, the effect of filter techniques on feature selection was analyzed in text datasets. For this purpose, various experiments were conducted on the six-text dataset frequently used in the literature. Three classification algorithms are considered in the experiments to analyze the classification and performance metrics in these datasets. While the results of six filter metrics were analyzed in the experiments, Precision, Sensitivity (Recall), Micro-F1, Macro-F1, and Accuracy metrics used in the literature for performance evaluation were examined. The novel Selection Error performance metric proposed in the study was compared with these performance metrics. As the experimental results show, reliance on classifiers in evaluating the success of a filter feature selection approach is not enough to see the actual performance of the approach. Overall, the DFS approach yields lower SE metric results in the experimental studies. In addition, the DFS approach has a more stable and balanced feature selection mechanism than other filtering methods.

This study has tried to find answers to some questions to determine which filter technique performs better. If another classifier was used instead of the NB classifier after the feature selection with the filter technique, and this classifier worked efficiently with one of the approaches other than DFS and IG, which approach would show the best performance? How would that be decided? In this study, the answer to this question was sought and tried to be answered. As can be seen in the experiments conducted as a result of the study, very successful results were obtained with the proposed SE metric. This study presents a novel performance metric to evaluate the performance of filter feature approaches. According to the results of many filter techniques on different classifiers and datasets in the experiments, the proposed novel performance metric is based on solid foundations.

In the literature, the performance of filtering techniques is evaluated together with classification methods. However, the performance metric proposed in this study is

classifier independent. When the results are analyzed, it is seen that there is a strong relationship between the proposed performance metric and the classification performance metric results. As can be seen from the results, the proposed novel performance metric will significantly contribute to the literature to see the effects of filter feature selection methods independent of classifier performance.

As with every metric, there are some limitations for the Selection Error. The change in the α threshold value used is the most significant limitation for the selection metric. Different outcomes can be obtained when the α value changes. The most ideal α mill in this case is the limitation of what is. Another limitation for SE is that in some data clusters, the distinguishing properties in the property space are difficult to determine. Furthermore, the negative region in the Selection Error is characterized by the fact that the distinction is not readily apparent. Additional operations may be required in these cases.

Conflict of Interest

The authors declare that we have no conflict of interest.

Acknowledgements

This work was supported by Siirt University, Fund of Scientific Research Projects under grant number 2020-SİÜMÜH-036

References

- [Aggarwal and Zhai, 2012] Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining text data*, 163-222.
- [Alberto et al., 2015] Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015). Tubespm: Comnt spam filtering on youtube. *IEEE 14th international conference on machine learning and applications (ICMLA)*, 138-143
- [Amazon Reviews, 2020] Amazon Reviews for Sentiment Analysis, (2020). Available from: <https://www.kaggle.com/datasets/bittlingmayer/amazonreviews>.
- [Asim et al., 2021] Asim, M., Javed, K., Rehman, A., & Babri, H. A. (2021). A new feature selection metric for text classification: eliminating the need for a separate pruning stage. *International Journal of Machine Learning and Cybernetics*, 12, 2461-2478.
- [Çekik and Uysal., 2020] Çekik, R., & Uysal, A. K. (2020). A novel filter feature selection method using rough set for short text data. *Expert Systems with Applications*, 160, 113691.
- [Çekik and Kaya, 2023]. Çekik, R., & Kaya, M. (2023). A New Feature Selection Metric Based on Rough Sets and Information Gain in Text Classification. *Gazi University Journal of Science Part A: Engineering and Innovation*, 10(4), 472-486.
- [Dumais et al., 1998] Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, 148-155.
- [Enron Email, 2015] Enron Email Dataset. (2015). Available from: <https://www.cs.cmu.edu/~enron/>.
- [Forman, 2003] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3, 1289-1305.

- [Go et al, 2009] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12).
- [Hancer et al, 2023] Hancer, E., Xue, B., & Zhang, M. (2023). An evolutionary filter approach to feature selection in classification for both single-and multi-objective scenarios. *Knowledge-Based Systems*, 280, 111008.
- [Jin et al., 2023] Jin, L., Zhang, L., & Zhao, L. (2023). Feature selection based on absolute deviation factor for text classification. *Information Processing & Management*, 60(3), 103251.
- [Kaya et al., 2013] Kaya, M., Bilge, H. Ş., & Yildiz, O. (2013). Feature selection and dimensionality reduction on gene expressions. 21st Signal Processing and Communications Applications Conference (SIU), 1-4.
- [Kaya and Bilge, 2016] Kaya, M., & Bilge, H. Ş. (2016). A hybrid feature selection approach based on statistical and wrapper methods. 24th Signal Processing and Communication Application Conference (SIU), 2101-2104.
- [Kou et al., 2020] Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86, 105836.
- [Labani et al., 2018] Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, 70, 25-37.
- [Labani et al., 2020] Labani, M., Moradi, P., & Jalili, M. (2020). A multi-objective genetic algorithm for text feature selection using the relative discriminative criterion. *Expert Systems with Applications*, 149, 113276.
- [Li et al., 2008] Li, Y., C. Luo, and S.M. Chung. (2008) Text clustering with feature selection by using statistical data. *IEEE Transactions on knowledge and Data Engineering*, 20(5), 641-652.
- [Liao and Vemuri, 2002] Liao, Y., & Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. *Computers & Security*, 21(5), 439-448.
- [Maimon and Rokach, 2014] Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications*, 81, World scientific.
- [McCallum and Nigam, 1998] McCallum, A., & Nigam, K. A. (1998). Comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, 1998, 41-48.
- [Nuruzzaman et al., 2011] Nuruzzaman, M. T., Lee, C., & Choi, D. (2011). Independent and personal SMS spam filtering. *IEEE 11th International Conference on Computer and Information Technology*, 429-435.
- [Ogura et al., 2009] Ogura, H., H. Amano, and M. Kondo. (2009). Feature selection with a measure of deviations from Poisson in text categorization. *Expert Systems with Applications*, 36(3), 6826-6832.
- [Parlak and Uysal, 2023] Parlak, B., & Uysal, A. K. (2023). A novel filter feature selection method for text classification: Extensive Feature Selector. *Journal of Information Science*, 49(1), 59-78.
- [Rehman et al., 2017] Rehman, A., K. Javed, and H.A. Babri. (2017). Feature selection based on a normalized difference measure for text classification. *Information Processing & Management*, 53(2), 2017, 473-489.

- [Rehman et al., 2018] Rehman, A., Javed, K., Babri, H. A., & Asim, M. N. (2018). Selection of the most relevant terms based on a max-min ratio metric for text classification. *Expert Systems with Applications*, 114, 78-96.
- [Rehman et al., 2015] Rehman, A., Javed, K., Babri, H. A., & Saeed, M. (2015). Relative discrimination criterion—A novel feature ranking method for text data. *Expert Systems with Applications*, 42(7), 3670-3681.
- [Reuters-21578, 1997] Reuters-21578 Text Categorization Collection Data Set. (1997). Available from: <https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection>.
- [Rish, 2001] Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 41-46.
- [Shang et al., 2007] Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert systems with applications*, 33(1), 1-5.
- [Scholkopf and Smola., 2018] Scholkopf, B., & Smola, A. J. (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.
- [Uysal and Gunal, 2012] Uysal, A.K. and S. Gunal. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, 36, 226-235.
- [Uysal et al., 2013] Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2013). The impact of feature extraction and selection on SMS spam filtering. *Elektronika ir Elektrotechnika*, 19(5), 67-72.
- [Yang et al., 1997] Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *International Conference on Machine Learning*, 412-420.
- [Yang, 1999] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2), 69-90.
- [Zhou et al., 2021] Zhou, H., Ma, Y., & Li, X. (2021). Feature selection based on term frequency deviation rate for text classification. *Applied Intelligence*, 51, 3255-3274.