Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 03
Issue 02

December, 2022

*Research* ***Article***

# Classification of Unwanted SMS Data (Spam) with Text Mining Techniques

***Rasim ÇEKİK***[1]

[1]*Department of Computer Engineering, Sirnak University,7300, Sirnak, Türkiye*

ARTICLE INFO

*Article history:*
Received November 26, 2022
Revised December 30, 2022
Accepted December 6, 2022

Keywords:
Feature Selection
Text Classification
Text Mining
Spam SMS

ABSTRACT

Text mining, which derives information from written sources such as websites, books, e-mails, articles, and online news, processes and structures data using advanced approaches. The vast majority of SMS (Short Message Service) messages are unwanted short text documents. Effectively classifying these documents will aid in the detection of spam. The study attempted to identify the most effective techniques on SMS data at each stage of text mining. Four of the most well-known feature selection approaches were used, each of which is one of these parameters. As a result, the strategy that yielded the best results was chosen. In addition, another parameter that produces the best results with this approach, the classifier, was determined. The DFS feature selection approach produced the best results with the SVM classifier, according to the experimental results. In Average Results, DFS showed the best result of 93.5361 for accuracy criterion, while it reached the highest result of 93.4953 for Macro-F1. This study establishes a general framework for future research in this area that will employ text mining techniques.

## 1. Introduction

With the rapid development of technology, the use of the Internet has increased tremendously. The influence of people on the internet has created a large amount of text documents. In this process, tools and techniques were needed to process text documents and provide access to information. The most effective method in this regard is text mining techniques, also called text analytics. Text Mining is the process of extracting previously unknown, potentially useful, structured and organized data from unstructured and disordered chunks of electronic text. Text mining, which infers from written sources such as websites, books, e-mails, articles, online news, processes the languages used in daily life and structures the data with the help of advanced approaches. Text mining tasks include text classification and clustering in general, concept or entity extraction, granular taxonomy modelling, sentiment analysis, entity relationship modeling, document abstraction, etc. transactions can be sorted.

In this study, the classification technique of text mining is used. Text classification is a process that includes preprocessing operations such as root finding, letter transformation, and techniques such as feature weighting and feature selection. Classification is the activity of categorizing unlabeled data according to the model created with the help of labeled texts. The most important problems encountered in the text classification stage are high dimensionality and space structure. Various techniques and approaches have been proposed to

[1] Corresponding author
e-mail: rasimcekik@sirnak.edu.tr
DOI: 10.55195/jscai.1210559

overcome these problems. Examples of this are the use of dimension reduction techniques against the high dimensionality problem or the use of feature selection approaches to select the sub-feature set that best expresses the entire feature space. Parlak B. and Uysal A. [1] proposed a new statistical feature selection approach as a solution to high dimensionality. The model utilized corpus-based and class-based probabilities for statistical calculations. Again, A. Uysal and S. Günal [2] presented the distinguishing feature selector (DFS), an effective and successful feature selection approach.

In its simplest definition, SMS (Short message service) can be defined as a short text message service used on phones. It is a popular application because it is a service used around the world. That's why it is used by so many people. This service, which is so widely used, may have some disadvantages. Just like in e-mails, the most annoying situation in the field of SMS is that unwanted SMS messages (spam) (eg credit announcements of banks, promotional messages of stores, discount announcements of e-commerce sites, tariff messages of mobile communication providers) fall into the message box of the mobile phone. This situation causes people to waste their precious time and fill the message box unnecessarily. Finding a solution to this problem or minimizing the effect of the problem is an important step for mobile users to prevent unnecessary use of both phone resources and time. The simplest method in this regard is the use of a black/white list, known as the banned list, where the phone numbers of the people who send the spam messages are added. This method is used in many commercial applications [3]. However, this method requires the intervention of the phone user and will block spam as well as regular messages from blacklisted phone numbers. This may result in missing important messages or loss of information, which may be beneficial to the phone user. For example, discount messages, coupons from a legitimate e-commerce platform, or favorable loan deals from a bank may be viewed as spam. More effective and efficient methods are needed to avoid similar situations. Therefore, it would be a more accurate approach to filter the texts according to the content of the incoming message rather than the phone number. Classifying texts according to their content is one of the main tasks of text mining. In this study, with the help of text mining techniques, it has been tried to classify whether SMS data is spam or not according to its content. Within the scope of the study, the most efficient framework model for SMS classification is revealed by using different text mining techniques and methods. Although there are a number of studies in the literature on SMS spam filtering, few studies have analyzed the effectiveness of the attributes used in the filtering process. In this study, it is examined that the classifiers achieve higher success by choosing features with high efficiency to classify SMS messages.

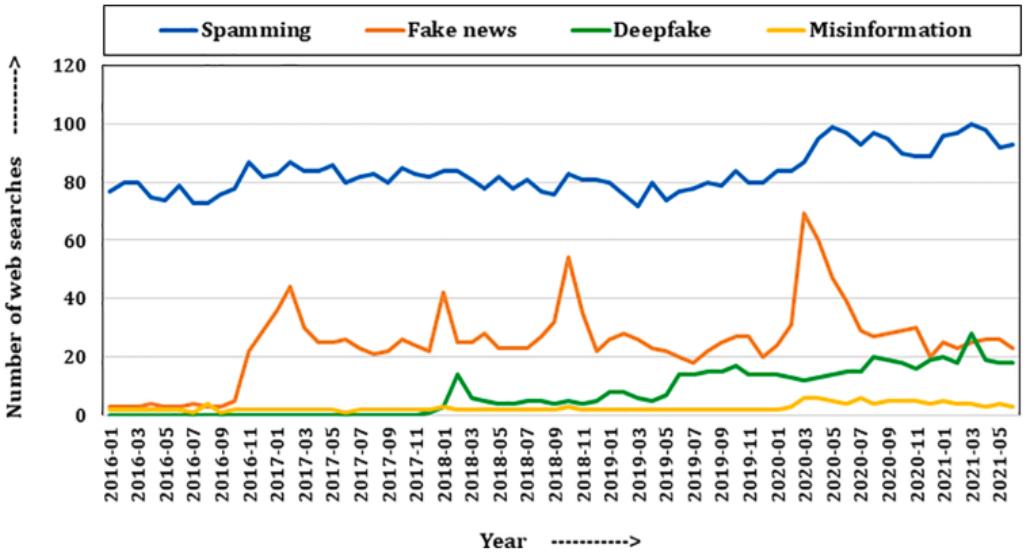
## 2. Literature Review

Telephone use has become one of the basic needs today. The phone is used in almost every part of life and people never leave it with them. Undoubtedly, one of the most frequently used applications in the vehicle, which has such an important use in our lives, is SMS. However, the vast majority of SMS data consists of spam messages. Rao S. et al. [17] showed in Figure 1 the Google trend analysis of web searches related to fake news, Deepfake and misinformation between 2016-2021. In addition, the fact that there are many studies in the literature on filtering the data of the SMS service shows the importance of the SMS application. For example, Delany S.J et al. [4] conducted a study covering a large literature review on SMS spam. In the study, motivating points about SMS spam were mentioned and some experimental studies were included. Similar studies have been done by research communities and have contributed to the field of spam detection [5-8]. It is possible to group the studies conducted in this field according to their methods. Studies were divided into 4 groups according to the methods used [9].

- machine learning,
- deep learning,
- spam graphic display
- spam filtering and detection with android application

Machine learning methods categorize messages as spam and non-spam with the help of classification and prediction approaches in data mining [10]. It is divided into two as labeled and unlabeled learning. Deep learning methods, on the other hand, try to detect spam using advanced versions of artificial neural networks [11]. Graphical representation

methods perform classification tasks using graphics in the text [12]. In the spam filtering and detection method with the Android application, a real-time mobile application is implemented to be run on mobile phones with Android operating system by using the feature vector and classifier pair that provides the most successful result in the simulation study [13]. In addition, the existing studies in the literature can be grouped under three headings: content-based, non-content-based and hybrid, according to the working principle of the approaches they use [14]. Content-based approaches [15] provide weighting of text documents using the bag of words (BoWs) method. With weighting, the frequency of occurrence of a word in the document can be determined. Non-content-based [16] approaches use message attributes or signature patterns as attributes to identify spam on the network. Hybrid approaches offer a model by combining the superior features of both content-based and non-content-based approaches.




Existing SMS filtering approaches can be structurally divided into three groups [12]. These are shown in Figure 2.

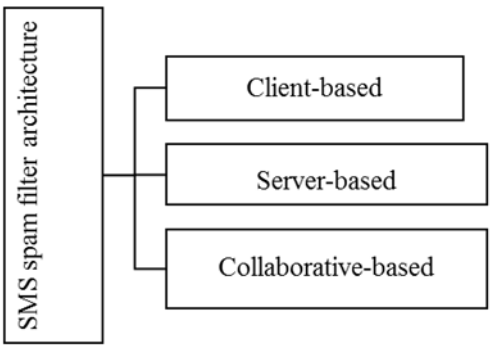



**Figure 2** Structurally approach groups for SMS filtering

- *Client-based:* Approaches that offer solutions by filtering on mobile devices.
- *Server-based:* Approaches that offer solutions by filtering on network providers.
- *Collaborative-based:* These are approaches that offer solutions by filtering on both network providers and mobile devices.

As the use of mobile devices in SMS filtering increases, the studies to be carried out in this area will remain up-to-date. Therefore, studies in this field continue without slowing down. Since this study is a content-based study, a summary of the studies on this subject is presented in Table 1.

**Table 1:** a summary of the content-based studies

| ref | Used machine learning approach |
|---|---|
| [18], [23] | k-NN classifiers |
| [19] | SVMs and others |
| [20] | Winnow algorithm |
| [21] | Bayes |
| [22] | Naive Bayes |

## 3. Text Mining and Techniques

Text Mining is also called Text Data Mining and Knowledge Discovery from Textual Databases. Although Text Mining is considered a sub-topic of

data mining, it is different from data mining. In Text Mining, it is the extraction of patterns from natural language texts rather than event-based databases. However, it offers knowledge discovery with stages similar to data mining. In summary, text mining is the process that aims structured data on unstructured text data. For example; It aims at studies such as classification of texts, clustering, extracting topics from texts, entity relationship modeling, sentiment analysis in texts, text summarization, author identification, and evaluation of customers' comments about the product. To achieve these goals, techniques such as text mining information extraction, information classification, syllable analysis, word frequency distribution, information extraction and visualization are used. The first of these methods is the text classification process.

Text classification offers a learning model by using available texts whose class is known beforehand. Then, the new incoming texts are classified according to this model. The biggest handicap in the classification process is the high dimensionality and the sparseness of the information system (IS). SMS the Information System is shown as $IS = (D, T)$. Where $D$ stands for finite non-empty universal text document set, $T$ stands for finite non-empty conditional and decision attribute set. $T = \{t, c\}$ specifies $t$ conditional, $c$ decision attributes. The processing order of the techniques used in the text classification process on the information system is given in Figure 4.

### 3.1. Feature Selection Approaches

Since the size of the feature space is very large, it is necessary to select the most representative subset of the entire feature space in order to perform an efficient classification in text mining. The most effective method for this process is to use feature selection approaches. Feature selection approaches are grouped under three headings [28, 29]: filter, wrapper, and embedded. The feature selection taxonomy is given in Figure 3.

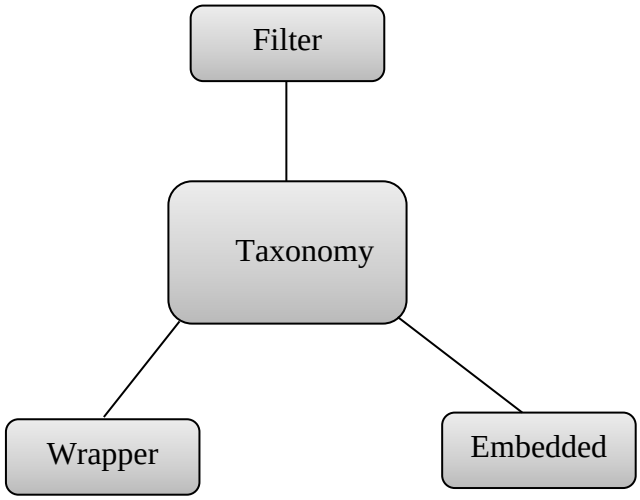



**Figure 3** The feature selection taxonomy

**Table 2** Brief information about used approaches

$$IG(t) = -\sum_{i=1}^{M} P(C_i) log P(C_i) + P(t) \sum_{i=1}^{M} P(t) log P(t) + P(\underline{t}) \sum_{i=1}^{M} P(\underline{t}) log P(\underline{t})$$

$P(C_i|t)$ and $P(C_i|\underline{t})$ show the conditional probability of class $C_i$ given presence and absence of the term $t$, respectively. $P(t)$ and $P(\underline{t})$ are the probabilities of absence and presence of the term $t$.

$$GI(t) = \sum_{i=1}^{M} P(C_i)^2 P(t)^2$$

The notation $P(C_i)$ in the formula indicates the probability that the $t$ term will be in the $C_i$ class.

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\underline{t}|C_i) + P\left(t|\underline{C_i}\right) + 1}$$

$P\left(t|\underline{C_i}\right)$ and $P(C_i|\underline{t})$ show the conditional probability of the term $t$ given classes other than $C_i$ and absence of the term t given class $C_i$ , respectively.

$$HI2(t,C) = \sum_{t\in\{0,1\}} \sum_{C\in\{0,1\}} \frac{(N_{t,C} - E_{t,C})^2}{E_{t,C}}$$

$$CHI2(t) = \sum_{i=1}^{M} P(C_i) * CHI2(t,C)$$

The N and E values represent the observed and expected frequency for each case of term t and class C, respectively.

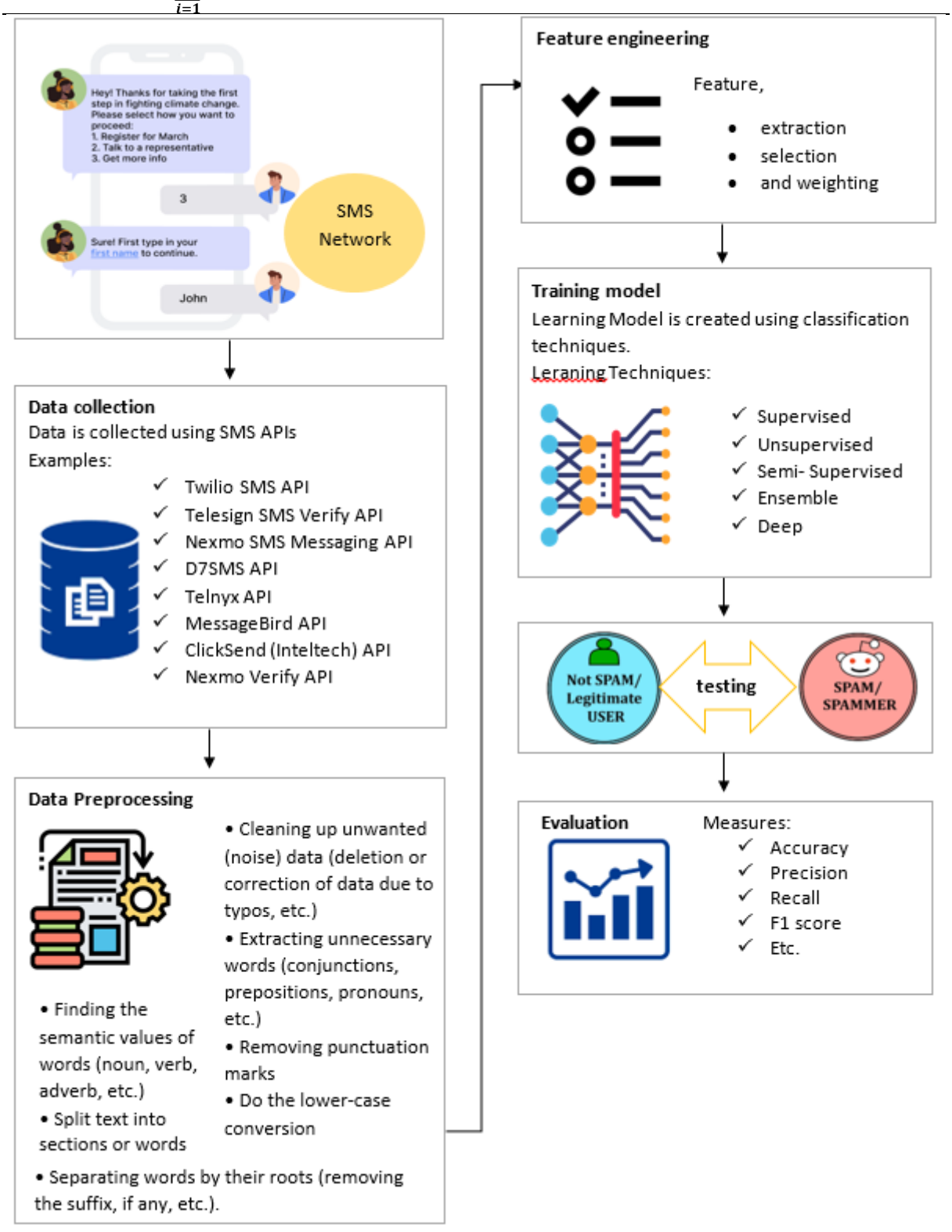



**Figure 4** SMS Filtering process

Filter approaches evaluate each attribute independently. Calculates a score for each attribute

with a statistical function. It selects N attributes according to this score. Such approaches generally work faster. Wrapper approaches select a subset of features using a classifier. Such techniques are based on the analysis of the relationship between feature subset selection and relevance. Embedded approaches create a subset of features by taking advantage of the best aspects of the other two feature selection approaches. Embedded techniques are based on independent criteria. These criteria are used to select optimal feature subsets with known cardinalities. While wrapper and embedded techniques require frequent classifier interaction during the feature selection phase, filter techniques do not. Three of the best-known approaches to fitters were used in this study: Information Gain (IG), Gini Index (GI), Distingushing Feature Selector (DFS) and Max-Min Ratio (MMR). Brief information about these approaches is given in Table 2.

### 3.2. Classification Methods

There are classifiers designed in various models in the literature [30]. The purpose of classifiers is to label unlabeled data using labeled data. In this section, the classifiers used in the study are mentioned.

#### *3.2.1. Support Vector Machines (SVM)*

SVM, which is based on the concept of margin maximization, is regarded as an effective classifier in the literature. Depending on the type of core used, it also has linear and non-linear versions. In this study, the linear version of SVM was used. The margin concept is central to the SVM classifier [24]. Classifiers separate classes using hyperplanes. Each hyperplane is distinguished by its direction (w) and precise position in space ($w_0$).

#### 3.2.2. K-Nearest Neighbors (KNN)

The k-nearest neighbors algorithm (KNN) is a non-parametric classification technique [25]. This technique is widely used in the field of text classification, as well as in many other fields. When a test document x is given, the approach finds k nearest neighbors of x among all documents in the training set and scores the category candidates based on the class of k candidates. The score of the category of neighboring documents can be the similarity of document x and each neighbor. For neighbor calculation, it employs one of the distance calculation methods such as Euclidean (Euclidean), Manhattan, and Minkowski. The approach takes the k-highest-scoring neighbor score after calculating the document's neighbor scores.

#### 3.2.3. *Naive Bayes (NB)*

In the field of text classification, the Naive Bayesian classifier has long been a popular method for categorizing texts. The theoretical foundation of the method is Thomas Bayes' theorem [26]. The independence of states in Bayes' theorem is the key to the naive Bayes classifier. As a result, the attributes are distinct. The classifier determines the probability of each situation and categorizes it based on the highest probability value.

## 4. Experimental Works

In this section, after giving information about the data set used in experimental studies, the success metrics used are introduced. Finally, the results obtained in the experimental studies are given and these results are analyzed.

### 4.1. Dataset

The SMS dataset is a collection of short text documents with free access on the Internet for detecting spam messages on phones. The dataset has been used in various studies to detect spam phone messages. First part of the collection consists of 450 manually retrieved messages on the Grumbletext Web site, where people in the UK are making public allegations about cell phone SMS spam messages. The second part of the collection is a subset of 3,375 SMS amateur messages randomly selected from the dataset of the NUS SMS Corpus (NSC), a dataset of approximately 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The third part of the collection is a list of 450 SMS amateur messages collected from Caroline Tag's PhD Thesis. Finally, in SMS Spam Corpus v.0.1 Big, 1.002 SMS amateur messages and 322 spam messages were added to the collection, resulting in the dataset known

as the SMS Spam Collection [27].

A subset of the SMS Spam collection was used in this study. 612 messages of the subset were used as training data and 263 as test data.

### 4.2. Success Criteria

- *Accuracy:* It comes from the beginning of the most frequently used success criteria in the classification process in data mining. The accuracy score is obtained by dividing the results of correctly classified samples by the total number of samples.

$$CS = size(\boldsymbol{Correctly\ Classified\ Samples})$$

$$\boldsymbol{AllS} = \boldsymbol{size}(\boldsymbol{Number\ of\ All\ Samples})$$

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{CS}}{\boldsymbol{AllS}}$$

- *Macro-F1 Criterion:* Macro f1, one of the F criteria, is a frequently used criterion in the text classification field. In the macro-average, equal weight is given to each class regardless of class frequency, since the mean of all classes is taken after the F measure is calculated for each class in the data set. Therefore, it is necessary to specify the Precision and Recall metrics before giving the Macro-F1 mathematical background. Macro-F1 obtained using these two metrics is calculated as follows.:

$$\boldsymbol{p_i} = \frac{\boldsymbol{1}}{\boldsymbol{C}} * \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{C}} \quad \boldsymbol{TP_i}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{C}} \quad \boldsymbol{TP_i + FP_i}}$$

$$\boldsymbol{r_i} = \frac{\boldsymbol{1}}{\boldsymbol{C}} * \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{C}} \quad \boldsymbol{TP_i}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{C}} \quad \boldsymbol{TP_i + FN_i}}$$

$$\boldsymbol{F_i} = \boldsymbol{2} * \frac{\boldsymbol{p_i * r_i}}{\boldsymbol{p_i + r_i}}$$

$$\boldsymbol{Makro - F1} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{C}} \quad \boldsymbol{F_i}}{\boldsymbol{C}},$$

In the formula, the pair $(p_i, r_i)$ corresponds to precision and precision for class $i$, respectively.

### 4.3.Accuracy Analysis

In experimental studies, random feature sizes Top-60, 110, 250, 450, 650, and 950 were chosen. The achievements of DFS, IG, G1, and CHI2 feature selection approaches for SVM, KNN, and NB classifiers in these feature Tops are given in Figures 5, 6, and 7. Figure 5, Figure 6, and Figure 7 show the results of the feature selection approaches for the SVM, KNN, and NB classifiers, respectively, in the aforementioned feature dimensions. Moreover, each figure represents results for both Accuracy and Macro F1.

When Figure 5 is examined, it can be observed that DFS offers the best results for the Accuracy criterion. Likewise, DFS showed the best results for Macro F1. Also, it can be seen in Figure 6 that IG performs better for both metrics when the KNN classifier is used. For the NB classifier, CHI2 and GI performed better. This situation is the same for both metrics (See Figure 7). Results in Figures show that each feature performs well with a classifier.

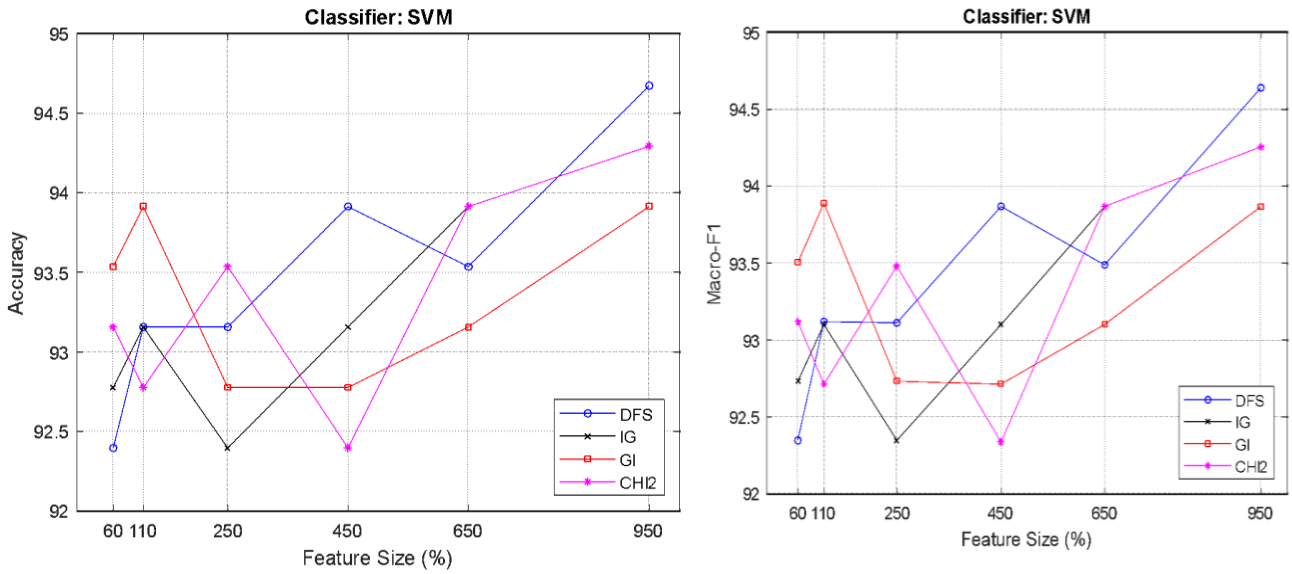



**Figure 5** Accuracy and Macro- F1 Results for SVM Classifier

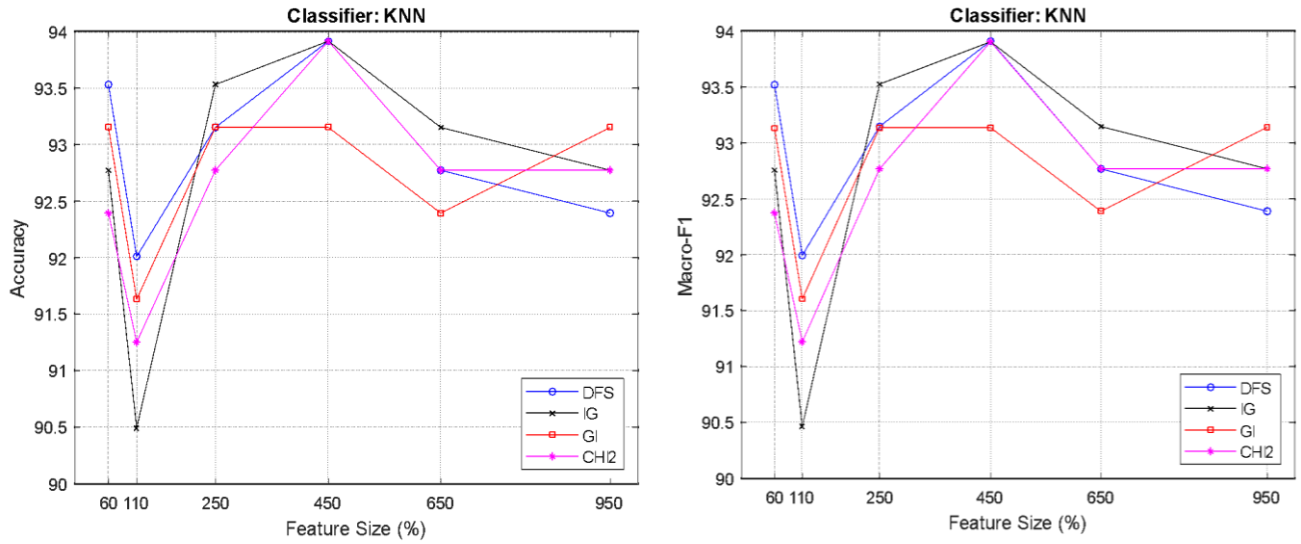



**Figure 6** Accuracy and Macro- F1 Results for KNN Classifier

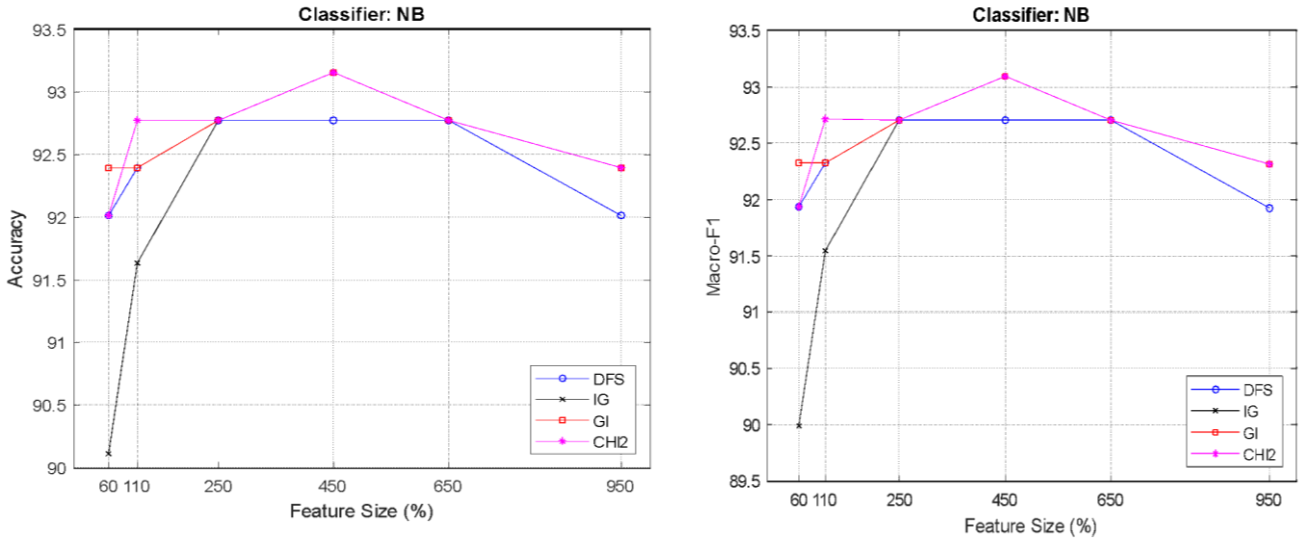



**Figure 7** Accuracy and Macro- F1 Results for NB Classifier

In order to better interpret the results, it is necessary to look at the average of the performances of each feature for all classifiers. Table 3 and Table 4 was created for this process.

**Table 3** Average Accuracy Results on all Classifiers

| Feature Size | DFS | IG | GI | CHI2 |
|---|---|---|---|---|
| 60 | 92.6489 | 91.8885 | 93.0291 | 92.5222 |
| 110 | 92.5222 | 91.7617 | 92.6489 | 92.2687 |
| 250 | 93.0292 | 92.9024 | 92.9024 | 93.0292 |
| 450 | **93.5361** | 93.4094 | 93.0292 | 93.1559 |
| 650 | 93.0292 | 93.2826 | 92.7757 | 93.1559 |
| 950 | 93.0291 | 93.1559 | 93.1559 | 93.1559 |

**Table 4** Average Macro-F1 Results on all Classifiers

| Feature Size | DFS | IG | GI | CHI2 |
|---|---|---|---|---|
| 60 | 92.6019 | 91.8289 | 92.9885 | 92.4777 |
| 110 | 92.4816 | 91.7054 | 92.6082 | 92.2174 |
| 250 | 92.9891 | 92.8604 | 92.8590 | 92.9860 |
| 450 | **93.4953** | 93.3686 | 92.9825 | 93.1132 |
| 650 | 92.9890 | 93.2416 | 92.7334 | 93.1153 |
| 950 | 92.9863 | 93.1420 | 93.1093 | 93.1142 |

When the tables are examined, DFS in 450 dimensions achieves the best performance for all classifiers based on both criteria. These tables are useful for determining whether feature selection approaches exhibit statistically stable performance. It's also useful to see which feature sizes perform best. For example, IG performed poorly in small dimensions compared to other approaches, but better in larger dimensions. Furthermore, DFS and CHI2 performed similarly and produced more consistent results than other approaches. This demonstrates that DFS and CHI2 collaborate closely. In other words, it chooses features that are close to being distinctive. These characteristics do not have to be the same. Assume that feature $A$ has a distinctness of 0.5689 and feature $B$ has a distinctness of 0.5772. In this case, DFS may select attribute $A$, while CHI2 may select attribute $B$. However, the contributions of these two features to the classifier are nearly equal.

As a result, the DFS approach appears to be more stable and performs better at certain feature Top-N dimensions. Furthermore, the GI approach performed well with small feature sizes. In terms of classifiers, it is appropriate to say that SVM outperforms others. These results show that using DFS and SVM classifiers results in more accurate SMS spam detection.

## 5. Conclusion

In this study, text mining techniques were used to analyze the SMS dataset to determine the necessary parameters for an effective classification study. Four of the most well-known feature selection approaches, each of which is one of these parameters, were employed. As a result, the approach that produced the best results was chosen. In addition, the classifier, another parameter that produces the best results with this approach, was determined. According to the experimental results, the DFS feature selection approach produced the best results with the SVM classifier. This study provides a general framework for future studies in this field that will use text mining techniques.

## References


[1]. Parlak, B., & Uysal, A. K. (2021). A novel filter feature selection method for text classification: Extensive Feature Selector. Journal of Information Science, 0165551521991037.

[2]. Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. Knowledge-Based Systems, 36, 226-235.

[3]. Android Apps. (Accessed March 2012). Available: https://play.google.com/store/apps

[4]. Delany, S. J., Buckley, M., & Greene, D. (2012). SMS spam filtering: Methods and data. Expert Systems with Applications, 39(10), 9899-9908.

[5]. Xiang, Y., Chowdhury, M., & Ali, S. (2004). Filtering mobile spam by support vector machine. In N. Debnath (Ed.), Proceedings of the third international conference on computer sciences, software engineering, information technology, E-business and applications (pp. 1–4).

[6]. Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Odusami, M. (2019). A review of soft techniques for SMS spam classification: Methods, approaches and applications. Engineering Applications of Artificial Intelligence, 86, 197-212.

[7]. Nagwani, N. K., & Sharaff, A. (2017). SMS spam filtering and thread identification using bi-level text classification and clustering techniques. Journal of Information Science, 43(1), 75-87.

[8]. Nagwani, N. K. (2017). A Bi-Level Text Classification Approach for SMS Spam Filtering and Identifying Priority Messages. International Arab Journal of Information Technology (IAJIT), 14(4).

[9]. Hanif, K., & Ghous, H. Detectıon Of Sms Spam And Fılterıng By Usıng Data Mınıng Methods: Lıterature Revıew.

[10]. Gupta, M., Bakliwal, A., Agarwal, S., & Mehndiratta, P. (2018). A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers. 2018 11th Internationalfile:///E:/Sms Spamming/Sms Spamming 15.Pdf Conference on Contemporary Computing, IC3 2018, 1–7. https://doi.org/10.1109/IC3.2018.8530469

[11]. Popovac, M., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2018). Convolutional Neural Network Based SMS Spam Detection. 2018 26th Telecommunications Forum, TELFOR 2018 - Proceedings, 1–4. https://doi.org/10.1109/TELFOR.2018.8611916

[12]. Lu Zhang, Jiandong Ding, Yi Xu, Yingyao Liu, and Shuigeng Zhou. 2021. Weakly-supervised text classification based on keyword graph. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages

2803–2813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[13]. Uysal, A. K., Günal, S., Ergin, S., & Günal, E. Ş. (2012, April). Detection of SMS spam messages on mobile phones. In 2012 20th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). Ieee.

[14]. Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Odusami, M. (2019). A review of soft techniques for SMS spam classification: Methods, approaches and applications. Engineering Applications of Artificial Intelligence, 86, 197-212."

[15]. Xiang, Y., Chowdhury, M., & Ali, S. (2004). Filtering mobile spam by support vector machine. In N. Debnath (Ed.), Proceedings of the third international conference on computer sciences, software engineering, information technology, E-business and applications (pp. 1–4).

[16]. Boykin, P. O., & Roychowdhury, V. P. (2005). Leveraging social networks to fight spam. IEEE Computer, 38, 61–68.

[17]. Rao, S., Verma, A. K., & Bhatia, T. (2021). A review on social spam detection: challenges, open issues, and future directions. Expert Systems with Applications, 186, 115742.

[18]. Healy, M., Delany, S., & Zamolotskikh, A. (2005). An assessment of case-based reasoning for short text message classification. In N. Creaney (Ed.), Proceedings of 16th Irish conference on artificial intelligence and cognitive science, (AICS-05) (pp. 257–266).

[19]. Gómez Hidalgo, J. M., Bringas, G. C., Sánz, E. P., & Garcı ´a, F. C. (2006). Content based SMS spam filtering. In D. Bulterman, & D.F. Brailsford (Eds.), Proceedings of the 2006 ACM symposium on document engineering DocEng '06 (pp. 107–114). New York, NY, USA: ACM.

[20]. Cai, J., Tang, Y., & Hu, R. (2008). Spam filter for short messages using winnow. In Proceedings of the international conference on advanced language processing and web information technology (pp. 454–459). IEEE.

[21]. Wu, N., Wu, M., & Chen, S. (2008). Real-time monitoring and filtering system for mobile SMS. In Proceedings of 3rd IEEE conference on industrial electronics and applications (pp. 1319–1324).

[22]. Jie, H., Bei, H., & Wenjing, P. (2010). A Bayesian approach for text filter on 3G network. In Proceedings of the 6th international conference on wireless communications networking and mobile computing (pp. 1–5).

[23]. Longzhen, D., An, L., & Longjun, H. (2009). A new spam short message classification. In Proceedings of the first international workshop on education technology and computer science (Vol. 2, pp. 168 –171).

[24]. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (s. (pp. 137-142). ). Berlin, Heidelberg.: Springer, .

[25]. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information.

[26]. Pearson, E. (1925). Bayes' theorem, examined in the light of experimental sampling. Biometrika.

[27]. https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

[28]. Cekik, R., & Uysal, A. K. (2020). A novel filter feature selection method using rough set for short text data. Expert Systems with Applications, 160, 113691.

[29]. Cekik, R., & Uysal, A. K. (2022). A new metric for feature selection on short text datasets. *Concurrency and Computation: Practice and Experience*, e6909.

[30]. Cekik, R., & Telceken, S. (2018). A new classification method based on rough sets theory. Soft Computing, 22(6), 1881-1889.

[31]. Parlak, B., & Uysal, A. K. (2021). The effects of globalisation techniques on feature selection for text classification. Journal of Information Science, 47(6), 727-739.