ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2021, 22(2), pp. 189-198, DOI: 10.18038/estubtda.898446

RESEARCH ARTICLE

A NEW MODEL ON AUTOMATIC TEXT SUMMARIZATION FOR TURKISH

Salih BAL [1,*], Efnan SORA GUNAL [2]

[1] Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Eskişehir Osmangazi Universiy, Eskisehir, Turkey
[2] Department of Computer Engineering, Faculty of Engineering and Architecture, Eskisehir Osmangazi University, Eskisehir, Turkey

ABSTRACT

The amount of data available in the electronic environment is increasing day by day with the development of technology. It becomes challenging and time-consuming for the users to access the information they desire within this increasing amount of data. Automatic text summarization systems have been developed to reach the desired information within texts in a shorter time than manual text summarization. In this paper, a new extractive text summarization model is proposed. In the proposed model, the inclusion of sentences of a given text in the summary is decided based on a classification approach. Also, the effectiveness of widely used features for automatic text summarization in the Turkish language is evaluated using sequential feature selection methods. The evaluations were carried out specifically for Turkish texts in the categories of economy, art, and sports. The experimental work justified the proposed text summarization method's performance and revealed how effective the features are.

Keywords: Text summarization, Feature selection, Classification

1. INTRODUCTION

Text summarization is the task of attaining a shorter version of one document or multiple documents in which the same main idea and prominent information are covered. If the summarization of a text is carried out by a computer, it is called automatic text summarization [1]. Nowadays, it is tough to acquire the brief data we want to obtain in the informational convergence. The study of summarizing one text or multiple texts by people to obtain essential data takes a long time and is difficult. Therefore, automatic text summarization systems are developed to reach the desired data quickly and easily.

Text summarization can be categorized as abstractive and extractive. In the abstractive summarization, the data in the original text is shortened and rewritten or paraphrased using linguistic features. As mentioned in [2], abstractive summarization is more complicated than extractive summarization and has not been reached a mature level yet. The same study also stated that none of the automatic text summarization systems is as successful as human assessors. On the other hand, in the extractive text summarization, the sentences chosen for the summary are not changed. Since the development of the system using this summarization method is less complex than abstractive summarization, many studies have preferred this approach rather than abstractive summarization [3-7].

The first study on text summarization in the English language [3] proposed a model that calculates the scores of sentences using term frequencies. Furthermore, the sentences which had high scores were parts of the summary in the proposed model. In [4], which is the first study in the Turkish language, term frequencies, and the location of sentences play prominent roles for text summarization. In [5], the authors proposed a text summarization system in the Indonesian language designed with a semantic

*Corresponding Author: salihbal26@gmail.com
Received: 17.03.2021 Published: 29.06.2021

analysis approach aiming to determine the similarity between sentences by using vector values of each sentence with the title. Further, in [6], the scores of sentences were used to determine which sentences were in summary with the help of a neural network. In that system, the features such as “title relevance,” “relative length of a sentence,” and “frequency of words,” which were extracted for the given input, were normalized from zero to one. These weights were used as the inputs for a feed-forward neural network. The output score, which shows the importance of the sentences, was obtained after implementing a feed-forward neural network. The sentence selection was performed by using these scores, and as a consequence, the summary was generated. Another study on extractive text summarization [7] focused on singular value decomposition. In this approach, Steinberger & Jezek and cross methods were used to select the sentences that are part of the summary. In [8], a text summarization system based on deep learning was implemented. The system was tested by using the Turkish dataset of news which has 13 different classes. The effects of the combinations of four criteria, which are "coverage," "redundancy-reduction," "relevance" and "coherence," on the performance of text summarization were examined in [9]. Eleven different combinations, including double, triple, and quadruple combinations, were tested in that study. It was observed that the "redundancy reduction" criterion affected the results the most, and the "coherence" criterion affected the result the least.

In line with the above-mentioned studies, a new extractive text summarization model is proposed in this paper. Also, the contributions of widely used features to automatic text summarization in the Turkish language is evaluated using sequential feature selection methods. The experimental studies were carried out using three different datasets. The first dataset consists of 100 texts in the categories of economy, arts, and sports. The second dataset [1, 10] consists of 20 texts in various categories: economy, art, health, and sports. The third dataset was produced by combining these two datasets.

The remaining of the paper is organized as follows: Section 2 describes the framework of the proposed model, including the dataset, preprocessing, sentence selection, feature selection, classification methods, and the new summarization model. The experimental work and results are discussed in Section 3. Finally, Section 4 presents the conclusions and possible future works.

2. FRAMEWORK OF THE PROPOSED MODEL

The framework of the system, which is shown in Figure 1, utilizes a Turkish dataset and includes preprocessing steps, features, feature selection methods, and classification algorithms. Each of these aspects is explained in the following subsections.

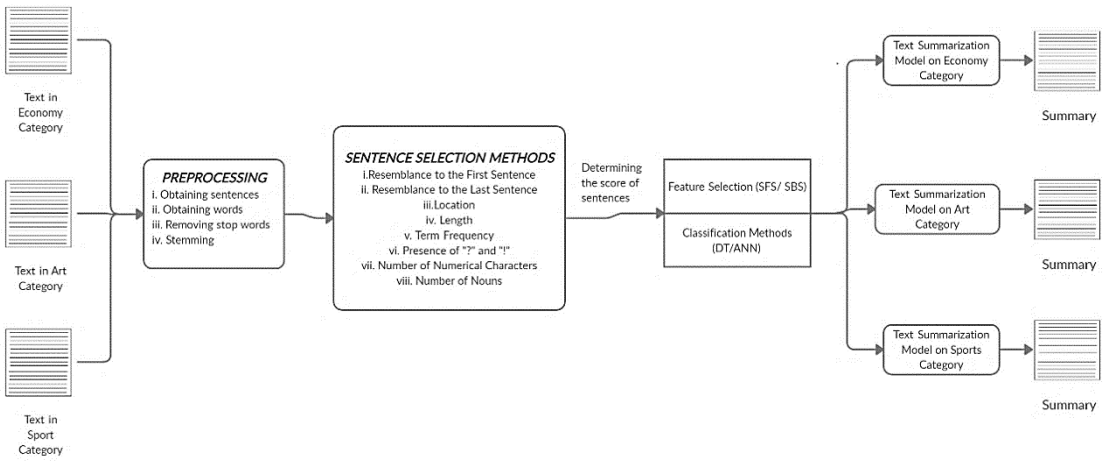

Figure 1. The framework of the proposed model

2.1. Datasets

Two datasets, namely Dataset I and Dataset II [1, 10], were used in Turkish news on the economy, sports, and art categories to investigate the contribution of eight different features to the success of text summarization. The documents in Datasets I and II were summarized by 5 and 30 different assessors, respectively. Also, a third dataset, namely Dataset III, is constituted by combining Datasets I and II. The attributes of these datasets are summarized in Table 1. During the experiments, no experiment was carried out specific to Dataset II due to its relatively small size. Instead, Dataset II was used to expand Dataset I and form Dataset III.

Table 1. The attributes of the datasets

Attribute	Dataset I	Dataset II	Dataset III
# Documents	100	20	120
# Sentences in Documents	1265	201	1466
Minimum Sentences/Document	4	7	4
Maximum Sentences/Document	42	10	42

2.2. Preprocessing

It has been proven that preprocessing improves the performance of automatic text summarization systems [5, 11, 12]. Therefore, in this stage, tokenization was applied to texts so that texts are split into sentences and sentences into words (or terms). Also, stop-word removal and stemming were applied as well. In the stemming step, Zemberek [13], a Turkish natural language processing library, was utilized. Following these preprocessing steps, the sentence selection methods, which are explained in the next subsection, were applied.

2.3. Features

In this paper, the contributions of eight widely-used features to automatic text summarization performance were evaluated. These features are summarized in Table 2. It should be also noted that all features were normalized to a range of [0,1] for further processing.

Table 2. The list of features

Feature No	Feature Name
f1	Resemblance to the First Sentence
f2	Resemblance to the Last Sentence
f3	Location
f4	Length
f5	Term Frequency
f6	Presence of "?" and "!"
f7	Number of Numerical Characters
f8	Number of Nouns

The descriptions of the features are as follows:

f1 - Resemblance to the First Sentence: The value of this feature for a given sentence is obtained based on the similarity of that sentence to the first sentence in the document. The similarity is calculated using the cosine similarity [14] as formulated in (1).

$$Score_{f1}(S_i) = \text{cosine}(S_i, S_{first}) \quad (1)$$

f2 - Resemblance to the Last Sentence: The value of this feature for a given sentence is obtained based on the similarity of that sentence to the last sentence in the document. The similarity is calculated using the cosine similarity [14] as formulated in (2).

$$Score_{f2}(S_i) = \text{cosine}(S_i, S_{last}) \quad (2)$$

f3 - Location: This feature was proposed by Edmunson [15] and applied based on the fact that certain sentences in the document have a higher probability of specifying the subject [14]. Equation (3) is used to compute the value of this feature for a sentence (S_i) in a given document, where TotNS is the total number of sentences and PS_i is the position of S_i in the document.

$$Score_{f3}(S_i) = (\text{TotNS} - \text{P}S_i)/\text{TotNS} \quad (3)$$

f4 - Length: The value of this feature for a given sentence corresponds to the number of words in that sentence [1].

f5 - Term Frequency: The value of this feature for a given sentence is calculated by adding the term frequencies of each term in the sentence [1, 16]. The term frequencies are computed considering the entire document that contains the sentence of interest. The term frequency is calculated as formulated in (4).

$$Score_{f5}(S_i) = \sum \text{Frequency of words in } S_i \quad (4)$$

f6 - Presence of "?" and "!": The value of this binary feature for a given sentence is 1 if the sentence contains a question mark or exclamation mark. Otherwise, it is 0 [14].

f7 – Number of Numerical Characters: The value of this feature for a given sentence corresponds to the number of numerical characters [14].

f8 – Number of Nouns: The value of this feature for a given sentence corresponds to the number of nouns in the sentence [1]. The nouns are detected using the Zemberek library [13].

2.4. Feature Selection

Feature selection methods are mainly divided into three categories, namely filter, wrapper, and embedded methods [17]. These methods are used in many fields, such as text mining [18-21], spam mail detection [22], and classification problems [23-25]. In filter methods, feature selection is independent of the classification algorithm. In wrapper methods, on the other hand, subsets of a full feature set are selected based on their performance in classification algorithms, where performance is usually measured using the classification accuracy or similar metric [26]. On the other hand, embedded methods contain both a classification algorithm and a feature selection method, classification, and feature selection operate simultaneously [26]. In this study, two different wrapper methods, including sequential forward selection (SFS) and sequential backward selection (SBS) were used to find the best subset of 8 features explained in the previous subsection. These suboptimal feature selection methods are widely used due to their speed and simplicity [27]. The search begins with an empty set for SFS. A feature is added to the selected subset at a time. In this way, the new subset maximizes the criterion function. This process is ended when this subset has the desired number of features. On the other hand, the search begins with all input features for SBS. A feature is eliminated from the feature set at a time. Therefore, the resultant subset maximizes the criterion function. This process is ended when the resultant feature set has the desired number of features.

2.5. Classification Algorithms

Text classification is one of the important and useful methods used in data mining and applied to various areas such as spam e-mail filtering, web page classification, and topic detection [28]. In our work, the decision tree (DT), which is a classifier algorithm in the form of a tree structure [29], and the artificial neural network (ANN), which is a computational model inspired by biological neural networks [30], were used as the classification algorithms due to their proven efficiency in various domains.

Decision trees contain special decision rules organized in a tree structure. In text classification, the document space is divided into non-overlapping areas at the decision tree's leaves. The prediction process is carried out on each leaf [29].

ANN is adapted from a simplified and concise view of neurons connected in layers to organize networks. ANN is used in many fields such as classification, control systems, and pattern recognition. ANN is a widely used method due to its fault tolerance, reliability, and learning ability [30].

2.6. The Proposed Summarization Model

A new extractive text summarization model is proposed in our work, where the summarization process is handled as a binary classification problem. For that purpose, the sentences in the datasets were labeled as “in-summary” and “out-summary”. For that purpose, it was determined how many assessors have labeled each sentence in the documents as a summarization sentence. If the majority of assessors have decided that the sentence will be in summary, this sentence of the document was labeled as "in-summary"; otherwise, it was labeled as “out-summary”. The sentence labeling algorithm is summarized below. The classification models are trained using the features extracted from the labeled sentences to determine whether a sentence will be in summary or out of summary.

Sentence Labeling Algorithm

- $D=\{S_i \mid i=1,2,\ldots,N\}$ represents the sentences in a given document where N is the total number of sentences.
- A is the total number of assessors.
- $vote_{i,j}$ represents the vote of A_j for S_i.
- The weights of each sentence are calculated as

  $$weightS_i = sum(vote_{i,j}),\ j=1,2,\ldots,A.$$

- Top-M sentences with the largest $weightS_i$ are labeled as “in-summary” where M is calculated as

  $$M=ceil(k \times N),$$

  where k defines the ratio of sentences used in the summary and defined to be 0.35 as indicated in [1].

- The remaining sentences are labeled as “out-summary”

3. EXPERIMENTAL WORK

During the experimental work, automatic text summarization is handled out using the proposed summarization model that is explained in the previous section. As mentioned earlier, the proposed model was evaluated on Datasets I, II, and III. For Dataset I, the documents from the categories of economy, art, and sports were used. For Datasets II and III, the documents from the categories of economy and sports were used. During the experiments, 10-fold cross-validation technique is used to

evaluate the results. The contribution of 8 features was determined using the DT and ANN classifiers together with the SFS and SBS feature selection methods.

Specifically, four groups of experiments were performed in the study. In the first and second experiments, the proposed classification-based summarization model was tested respectively on Dataset I and III. In the third and fourth experiments, the summarization model proposed in [1, 10] was employed using only the best feature subsets selected in the former experiments rather than the full feature set.

The results of the first two experiments are presented in Tables 3 and 4, respectively, where the selected features are marked, and the highest accuracies are indicated in bold.

Table 3. The results for Dataset I in the category of (a) economy (b) art (c) sports.

Feature Selection	ALL		SFS		SBS	
Classifier	DT	ANN	DT	ANN	DT	ANN
Accuracy	76.05	77.84	79.24	**80.84**	78.44	80.64
f1	✓	✓		✓	✓	
f2	✓	✓		✓	✓	✓
f3	✓	✓	✓	✓		✓
f4	✓	✓				✓
f5	✓	✓				
f6	✓	✓				
f7	✓	✓		✓	✓	✓
f8	✓	✓				

(a)

Feature Selection	ALL		SFS		SBS	
Classifier	DT	ANN	DT	ANN	DT	ANN
Accuracy	74.10	72.18	**79.34**	78.79	77.14	78.51
f1	✓	✓	✓	✓	✓	✓
f2	✓	✓	✓	✓		
f3	✓	✓	✓			✓
f4	✓	✓				
f5	✓	✓				
f6	✓	✓				
f7	✓	✓		✓		
f8	✓	✓			✓	

(b)

Feature Selection	ALL		SFS		SBS	
Classifier	DT	ANN	DT	ANN	DT	ANN
Accuracy	71.57	68.33	72.82	**74.81**	72.57	74.06
f1	✓	✓		✓	✓	✓
f2	✓	✓	✓		✓	
f3	✓	✓	✓	✓	✓	✓
f4	✓	✓	✓	✓		✓
f5	✓	✓			✓	✓
f6	✓	✓	✓			
f7	✓	✓				✓
f8	✓	✓	✓			

(c)

Table 4. The results for Dataset III in the category of (a) economy (b) sports.

Feature Selection	ALL		SFS		SBS	
Classifier	DT	ANN	DT	ANN	DT	ANN
Accuracy	75.70	75.70	77.02	78.18	77.02	**79.67**
f1	✓	✓	✓	✓	✓	
f2	✓	✓				✓
f3	✓	✓	✓	✓	✓	✓
f4	✓	✓	✓		✓	✓
f5	✓	✓				
f6	✓	✓		✓		
f7	✓	✓		✓	✓	✓
f8	✓	✓	✓			

(a)

Feature Selection	ALL		SFS		SBS	
Classifier	DT	ANN	DT	ANN	DT	ANN
Accuracy	70.06	66.67	71.95	**74.73**	71.97	73.04
f1	✓	✓	✓		✓	✓
f2	✓	✓		✓		✓
f3	✓	✓		✓		
f4	✓	✓				✓
f5	✓	✓				
f6	✓	✓				✓
f7	✓	✓				
f8	✓	✓				✓

(b)

When the classification results obtained in all three categories are considered, it is seen that using appropriate feature subsets rather than using all features provides higher classification performance. Considering the feature selection methods, it was observed that the features selected by the SFS method are more effective than those selected by the SBS method. Also, the features "Resemblance to the First Sentence (f1)" and "Location (f3)" were used in common to obtain the highest summarization performance for all categories in Dataset I.

To compare the performance of the proposed method, documents in the studies [1, 10] were used. As shown in Table 4, the features "Resemblance to the Last Sentence (f2)" and "Location (f3)" were used in common to obtain the highest summarization performance for all categories in Dataset III.

For the first two experiments, the experimental results listed in Tables 3 and 4 were obtained by the proposed classification-based summarization model. On the other hand, for the third and fourth experiments, the score values (f1, f2, f3, f4, f5, f6, f7, f8) obtained from the used features are summed for each sentence in the text. Achieved scores of the sentences are listed in ascending order. 35% of them with the highest score are labeled as "in-summary," and the others are tagged as an "out-summary" sentence. The success rate (SR) employed in [1, 10] was used for a fair comparison as formulated in (5).

$$\text{Success Rate} = \frac{|S \cap T|}{S} \tag{5}$$

In this equation, S is the number of sentences that are selected to be in the summary by the summarization model, and T is the number of sentences that are selected to be in the summary by the assessors. The experimental results obtained in the third and fourth experiments together with the best results obtained in the first two experiments are comparatively presented in Table 5.

Table 5. Comparison of the results of the four experiments.

Experiment #	Dataset	Category	Summarization Model	Feature Selection	Accuracy/SR (%) (with the Selected Features)	# Selected Features	Accuracy/SR (%) (with All Features)
1	I	Economy	ANN	SFS	80.84	4 (f1,f2,f3,f7)	77.84 (ANN)
2	III	Economy	ANN	SBS	79.67	4 (f2,f3,f4,f7)	75.70 (DT/ANN)
3	I	Economy	Eq.(5)	-	50.23	4 (f1,f2,f3,f7)	-
4	III	Economy	Eq.(5)	-	55.45	4 (f2,f3,f4,f7)	-
1	I	Art	DT	SFS	79.34	3 (f1,f2,f3)	74.10 (DT)
3	I	Art	Eq.(5)	-	59.46	3 (f1,f2,f3)	-
1	I	Sports	ANN	SFS	74.81	3 (f1,f3,f4)	71.57 (DT)
2	III	Sports	ANN	SFS	74.73	2 (f2,f3)	70.06 (DT)
3	I	Sports	Eq.(5)	-	58.64	3 (f1,f3,f4)	-
4	III	Sports	Eq.(5)	-	54.67	2 (f2,f3)	-

As shown in Table 5, it is observed that using ANN classifier and SFS feature selection method in the economy and sports categories for Dataset I, the accuracy is improved by around 3%. The feature "Location (f3)" is available in all of the most successful feature combinations. Besides, the feature "Number of Numerical Characters (f7)" plays an active role in these combinations. Also, the ANN classifier is the most successful one in the economy and sports categories. However, the highest accuracy is attained using the DT classifier in the art category. Moreover, among eight features, it is apparent that there are no contributions of "Term Frequency(f5)", "Presence of "?/!"(f6)" and "Number of Nouns (f8)" to the success of text summarization.

4. CONCLUSIONS

In this paper, a new extractive text summarization model is explicitly proposed for Turkish texts. In the proposed model, the inclusion of sentences of a given text in summary is decided based on a classification approach. Also, the contributions of widely used features to summarization performance are evaluated using feature selection methods. The evaluations were carried out for different datasets, including news texts on the economy, art, and sports. The results of the experimental work show that text summarization performance can be improved using appropriate features. The experiments also verified the performance of the proposed model and revealed the effectiveness of the features utilized. The evaluation of different features and text topics as well as performing topic classification before text summarization remain interesting future works.

CONFLICT OF INTEREST

The authors stated that there are no conflicts of interest regarding the publication of this article.

REFERENCES

[1] Güran A, Arslan SN, Kılıç E, Diri B. Sentence selection methods for text summarization. In: 2014 22nd Signal Processing and Communications Applications Conference (SIU), IEEE, 192-195.

[2] alZahir S, Fatima Q, Cenek M. New graph-based text summarization method. In: 2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 396-401.

[3] Luhn HP. The Automatic Creation of Literature Abstracts. IBM J. Res. Dev., 2, 159-165.

[4] Altan Z. A Turkish Automatic Text Summarization System. In:2004 IASTED International Conference on AIA, 16-18 February.

[5] Tardan PP, Erwin A, Eng KI, Muliady W. Automatic text summarization based on semantic analysis approach for documents in Indonesian language. In: 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), 47-52.

[6] Hingu D, Shah D, Udmale S. Automatic text summarization of Wikipedia articles. In: 2015 International Conference on Communication, Information & Computing Technology (ICCICT), 1-4.

[7] GeethaJ K, Deepamala N. Kannada text summarization using Latent Semantic Analysis. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1508-1512.

[8] Karakoç E, Yılmaz B. Derin Öğrenme Tabanlı Yoruma Dayalı Türkçe Haber Özetleme. In: 2019 22nd Signal Processing and Communications Applications Conference (SIU), IEEE.

[9] Sanchez-Gomez JM, Vega-Rodríguez MA, Perez CJ. Experimental analysis of multiple criteria for extractive multi-document text summarization. Expert Systems with Applications. 2020; 140, 112904. https://doi.org/10.1016/j.eswa.2019.112904.

[10] Güran A, Uysal M, Ekinci Y, Güran CB. An Additive FAHP based sentence score function for text summarization. Inf. Technol. Control. 2017; 46(1), 53-69.

[11] Gulati AN, Sawarkar S. A novel technique for multidocument Hindi text summarization. In: 2017 International Conference on Nascent Technologies in Engineering (ICNTE), 1-6.

[12] Yadav J, Meena YK. Use of fuzzy logic and wordnet for improving performance of extractive automatic text summarization. In : 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (2016, September), IEEE, 2071-2077.

[13] Akin AA, Akin MD. Zemberek, an open source NLP framework for Turkic Languages. Structure 10: 1-5, 2007.

[14] Güran A. Otomatik Metin Özetleme Sistemi, Phd. Thesis, Yildiz Technical University, Istanbul, 2013.

[15] Edmundson HP. New methods in automatic extracting. Journal of the Association for Computing Machinery. 1969; 16: 264-285.

[16] Güran A, Bayazıt NG, Gürbüz MZ. Effcient feature integration with Wikipedia-based semantic feature extraction for Turkish text summarization. Turkish Journal of Electrical Engineering & Computer Sciences. 2013; 21(5):1411-1425.

[17] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bionformatics. Bionformatics. 2007; 23: 2507-2517.

[18] Forman G. An Extensive Emprical Study of Feature Selection Metrics for Text Cassification. Journal of Machine Learning Research 2003; 3: 1289-1305.

[19] Liu T, Liu S, Chen Z, Ma WY. An evaluation on feature selection for text clustering. In: 2003 Proc. of the 20th international conference on machine learning (ICML-03), 488-495.

[20] Bai X, Gao X, Xue B. Particle swarm optimization based two-stage feature selection in text mining. In: 2018 IEEE Congress on Evolutionary Computation (CEC) (2018, July), 1-8.

[21] Mihuandayani Utami E, Luthfi ET. Text mining based on tax comments as big data analysis using SVM and feature selection. In: 2018 International Conference on Information and Communications Technology (ICOIACT), 537-542.

[22] Mohamad M, Selamat A. An evaluation on the efficiency of hybrid feature selection in spam email classification. In: 2015 International Conference on Computer, Communications, and Control Technology (I4CT), 227-231.

[23] Yildiz O, Tez M, Bilge HŞ, Akcayol MA, Güler I. Meme kanseri sınıflandırması için gen seçimi. In: IEEE 20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, 2012.

[24] Canedo VB, Marono NS, Betanzos AA, Benitez JM, Herrera F. A review of microarray datasets and applied feature selection methods. Information Sciences. 2014; 282: 111-135.

[25] Onan A. Ensemble learning based feature selection with an application to text classification. In: IEEE 2018 26th Signal Processing and Communications Applications Conference (SIU).

[26] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003; 3(Mar): 1157-1182.

[27] Nakariyakul S, Casasent DP. An improvement on floating search algorithms for feature subset selection. Pattern Recognition. 2009; 42, No. 9: 1932-1940.

[28] Uysal AK, Günal S. The impact of preprocessing on text classification. Information Processing and Management. 2014; 50: 104-112.

[29] Weiss SM, Indurkhya N, Zhang T. Fundamentals of predictive text mining. Springer, 2015.

[30] Suzuki K. Artificial Neural Networks – Architectures And Applications. 2013, Intech open science, ISBN 978-953-51-0935-8.