Journal of Soft Computing and Artificial Intelligence

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

International Open Access

Volume 04
Issue 01

June, 2023

Research ***Article***

Comparative Analysis of Globalisation Techniques for Medical Document Classification

Bekir Parlak[1], ***Salih Berkan Aydemir***[1]

[1]*Department of Computer Engineering, Amasya University, 0500, Amasya, Turkey*

ARTICLE INFO

Article history:
Received **December 9, 2022**
Revised **February 13, 2023**
Accepted **February 17, 2023**

Keywords:
Medical documents
Text Classification
Feature selection
Globalisation techniques

ABSTRACT

Medical document classification is one of the important topics of text mining. Globalisation techniques play a major role in feature selection stage. Therefore, globalization techniques affect text classification performance. Our aim in the study is to conduct a detailed analysis on two data sets with English and Turkish content by using medical text summaries of Turkish articles. These datasets consist of Turkish and English text summaries of the same articles. To observe how successful local feature selection methods in the field of text classification affect the classification performance on these two equivalent data sets by applying different globalisation techniques. The feature selection methods used are CHI2(chi-square), MI (mutual information), OR (odds ratio), WLLR (weighted log-likelihood ratio). Globalisation techniques are SUM (summation), AVG (average), MAX (maximum). Classifiers are MNB (multinomial naive bayes), DT (decision tree), and SVM (support vector machine). For the English Ohsumed data set, the highest Micro F score value of 95.48 was obtained in the max globalization method with the 2000-dimension CHI2 feature selection method and MNB classifier method. For the Turkish Ohsumed data set, the highest Micro F score value of 92.75 was obtained in the max globalization method with the 2000-dimension CHI2 feature selection method and MNB classifier method. In comparisons, it has been observed that the best classifier for Ohsumed datasets is MNB.

1. Introduction

With the rapid development of internet technologies in recent years, it can be seen that there is a huge increase in the number of electronic documents. The fact that the internet is more accessible to people and the increase in personal computers are among the reasons for this increase. Text classification methods play an important role in many documents on the internet. It can be used in solving various problems such as text classification [1], spam filtering [2], author identification [3], classification of web pages [4], classification of medical texts [5]. The importance of text classification increases the importance of databases where text classification is used. Documents in the database named MEDLINE are generally used for information access to medical texts and text classification studies. MEDLINE is a bibliographic database containing over 21 million documents from approximately 5.600 medical journals. This database can be queried with certain parameters over the internet, thanks to a search platform called PubMed [6]. In Turkey, there is TUBITAK's Medical Database created to facilitate access to information for experts working in the field of medicine. MEDLINE and ULAKBIM Medical Database are indexed by taking the relevant MeSH (Medical Subject Headings) terms with category information and selecting them manually by experts. Although an automated system is not used for indexing the

[1] Corresponding author
e-mail: salih.aydemir@amasya.edu.tr
DOI: 10.55195/jscai.1216800

MEDLINE database, there are automatic text classification studies on MEDLINE data in the literature. On the other hand, more useful and concise data are obtained by applying various methods on medical document data. These methods are featuring weighting, classification, feature selection, pre-processing and feature extraction. The text properties corresponding to a large number of documents are also quite high. Therefore, the dimensionality disaster will be greatly affected if size reduction is not made in the face of high dimensional text features. Feature extraction and feature weighting are the two main methods of reducing feature dimensionality. Feature weighting is a text classification phase that calculates the feature weight for each feature of documents. Feature extraction is a size reduction process in which an initial dataset is reduced to more manageable groups for processing.

In text categorization (TC), feature selection can be applied after feature extraction. Considering local feature selection methods, a globalisation policy is required to transform multiple local scores into a unique global score [7]. Globalisation techniques play an important role in TC. Considering the local scores, the global score can be calculated using various globalisation techniques. On the other hand, pre-processing is an important step for TC. Here are some pre-processing methods for text classification: lowercase conversion, removal of stop words, stemming and tokenization [8].

The motivation of this study is to choose the ideal feature selection, classifier and globalization techniques for Turkish and English Ohsumed datasets. All experiments were repeated in different dimensions, Macro F1 and Micro F1 values were calculated, and the results were reported.

Other parts of the work are organized as follows. In the second part, a detailed study area is examined. The basic methods used in the study are explained in the third part. In fourth part, experimental studies are given. In the last part, the conclusion part and future studies are mentioned.

2. Literature Review

Until now, various feature selection methods and classifiers have been applied on TC. In this part, feature selection methods used in TC are included.

Zheng et al. used information gain (IG), chi-square (CHI), correlation coefficient (CC) and odds ratios (OR) feature selection methods on imbalance data. Authors discussed feature selection methods in both one-sided (CC, OR) and two-sided (IG, CHI) metrics [9]. SVM produces effective results for TC. Taire and Haruno have investigated the effect of prior feature selection for Support Vector Machine (SVM) [10]. There are new feature selection methods proposed for TC in the literature. Gunal has proposed a novel hybrid feature selection which combine filter and wrapper methods for text classification [11]. In another study, Biricik et al. proposed a supervised feature extraction algorithm by combining the effect of input properties on classes. Their method is called abstract feature extraction [12].

Conventional TC algorithms consist of three main parts as handcrafted, nature-inspired and graph-based [13]. In the field of TC, many optimization-based feature selection methods have been proposed. The sine-cosine optimization algorithm, which has been proposed inspired by the sin and cos curves, has been developed and it is used as the feature selection method [14]. Feature selection was proposed using PSO. In addition, radial basis function neural networks are used as classifiers [15]. On the other hand, TC is applied using handcrafted features [13]. Some scholars have used traditional classifiers for the creation of feature sets and classification purposes, and they have proposed graph based feature selection methods [16].

Although feature selection and classification algorithms play an active role in a TC problem, globalisation techniques have strong effects on TCs. Some of the feature weighting methods in the literature generate a single global weighting score for each feature. However, local-based methods produce a different score for each class. There are some ways to get global scores from local scores: maximization, average, weighted average, and weighted maximum are popular globalisation techniques [17]. Parlak and Uysal have performed the impact of globalisation techniques on feature selection methods in TC [5]. In their studies, they used two successful classifiers, while they used four benchmark data sets. For Turkish Ohsumed dataset, the highest Micro-F1 and Macro-F1 scores are 92.75 and 82.82, respectively. It was obtained with the combination of CHI2 method, MAX globalisation technique, and MNB classifier using 2000 feature size. SVM classifier is the successor classifier for most cases. Also, CHI2 method is more successful than the other feature selection method in most cases for this data set. MI is the worst feature selection method for all situations.

3. Preliminaries

In this section, the basic techniques used in the study are mentioned.

3.1 Classifiers

The aim of text classification studies is to classifying uncategorized documents into predefined classes. In our experiments, three successful classifiers were employed to evaluate selected features by different globalisation techniques for each dataset. These classifiers are Multinational Naive Bayes (MNB), Decision Tree (DT), and Support Vector Machines (SVM). MNB is a form of naive Bayes classifier and very successful classifiers in text classification domain [37]. As classic Naive Bayes models a document with the occurrence and not occurrence of certain features, MNB clearly models it using feature counts. Multinational and multi-variate Bernoulli event models are widely utilized for text classification studies. While MNB takes into account term frequencies, multi-variate Bernoulli event model employs document frequencies. DT is one of the most efficient classifiers in text classification domain [38]. DT is a nonlinear classifier where classes are not accepted until a logical class is detected. SVM is one of the best classifiers in text classification studies. It has two versions which are linear and non-linear. In the experiments, we employed linear version of SVM classifier. The main subject of SVM classifier is the margin. LibSVM library is used for SVM classifier with linear kernel [39].

3.2. Feature Selection Methods

In our experiments, we employed four local feature selection algorithms. These are Chi-Square (CHI2), Mutual Information (MI), Odds Ratio (OR), and (WLLR).

CHI2: CHI2 is a successful feature selection method in text classification domain. The CHI2 method calculates the lack of independence between feature t and class C [17]. A and B events are assumed to be independent if

$$p(XY) = p(X)p(Y) \tag{1}$$

CHI2 method can be calculated as below:

$$CHI2(t_i, c_i) = \frac{N*(TP*TN-FP*FN)^2}{(TP+FN)*(TP+FP)*(FN+TN)*(FP+TN)} \tag{2}$$

MI: MI is a local method which computes the correlation between classes and features [18]. MI is computed as below:

$$MI(t_i, c_j) = log\frac{P(t_i|c_j)}{P(t_i)} \tag{3}$$

OR: OR is a supervised and local feature selection method which calculates the membership and non-membership to each class by utilizing nominator and denominator in Equation 4, respectively [19]. So, the OR method can produce both the negative and the positive scores. The method is computed as:

$$\boldsymbol{OR(t_i, c_j) = log\frac{P(t_i|c_j)(1-P(t_i|\bar{c}_j))}{(1-P(t_i|c_j))P(t_i|\bar{c}_j)}} \tag{4}$$

WLLR: WLLR is a supervised and local feature selection method which is proposed by Nigam et al. [20]. The WLLR method is calculated as below:

$$WLLR(t_i, c_j) = P(t_i|c_j)log\frac{P(t_i|c_j)}{P(t_i|\bar{c}_j)} \tag{5}$$

3.3. Globalisation Techniques

In our experiments, MAX, SUM, AVG globalisation techniques were utilized [5]. The reason we use these methods is to examine in detail how the same feature selection methods affect performance with different globalization techniques. These methods are generally used in the literature.

All of the scores are summed in SUM technique. The scores computed on each class are globalized by multiplying class probabilities in AVG technique. In MAX technique, the maximum of all scores is taken. Here, $f(t_i, C_j)$ corresponds to the score of the feature t_i in class C_j. These globalization techniques can be calculated as below:

$$SUM = \sum_{j=1}^{M} f(t_i, c_j) \tag{6}$$

$$AVG = \sum_{j=1}^{M} P(C_j) * f(t_i, c_j) \tag{7}$$

$$MAX = \max_{j=1}^{M} f(t_i, c_j) \tag{8}$$

4. Experimental Study

In our experiments, we used two data sets. Micro-F1 and Macro-F1 scores were utilized to analysing classification performance. Ten largest classes were included in the experimental works. The characteristics of the data sets used in the article are given in Table 1 and Table 2. Within the scope of the

study, experiments were carried out using java programming and WEKA tool. The flow chart of the analyzes made in Figure 1 has been added.

10-fold cross-validation was employed for fair evaluation. Different number of features which were selected by each feature selection method were fed into MNB, SVM and DT classifiers. 100, 250, 500, 1000 and 2000 dimension was used as a feature size. Also, the total number of features are 8610, 14334 for English and Turkish data sets, respectively. Resulting Micro-F1 and Macro-F1 scores are showed in Tables 3-8. For English Ohsumed dataset, the highest Micro-F1 and Macro-F1 scores are 95.48 and 88.25, respectively.

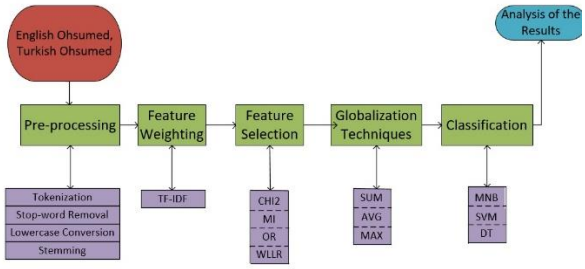

Figure 1 Flowchart of applied analyzes.

Table 1 Ohsumed Dataset for English

Class Number	Disesase Category	*Number of Documents*
1	Bacterial Infections and Mycoses	631
2	Virus Diseases	249
3	Parasitic Diseases	183
4	Neoplasms	2513
5	Musculoskeletal Diseases	505
7	Stomatognathic Diseases	132
8	Respiratory Tract Diseases	634
10	Nervous System Diseases	1328
14	Female Genital Diseases and Pregnancy Complications	2876
23	Pathological Conditions, Signs and Symptoms	1924

Table 2 Ohsumed Dataset for Turkish

Class Number	Disesase Category	*Number of Documents*
1	Bacterial Infections and Mycoses	284
2	Virus Diseases	44
3	Parasitic Diseases	116
4	Neoplasms	32
5	Musculoskeletal Diseases	140
7	Stomatognathic Diseases	39
8	Respiratory Tract Diseases	90
10	Nervous System Diseases	83
14	Female Genital Diseases and Pregnancy Complications	231
23	Pathological Conditions, Signs and Symptoms	73

The best score was obtained from the combination of CHI2 method, MAX method and MNB classifier using 2000 feature size. DT classifier is the second successful classifier. Also, CHI2 method is more successful than the other feature selection method in most cases for this dataset. MI is the worst feature selection method for all situations. Also, MAX globalisation is more efficient method than the other globalisation according to each feature selection method for most cases. For english dataset, in Table 3, in the classifications made with MNB, the best values were obtained with CHI2. In Table 4, in the classifications made with SVM, the best values were obtained with CHI2 in AVG. In Table 5, in the classifications made with DT, the score values were obtained with CHI. For turkish dataset, In Table 6, in the classifications made with MNB, the score values were obtained with CHI2 in MAX. In Table 7, in the classifications made with SVM, the best values were obtained with WLLR in AVG. In Table 8, in the classifications made with DT, the best values were obtained with OR in MAX.

Generally speaking, the performance increases as the number of dimesions increases in the datasets. While the highest scores in the English data set are obtained with the CHI2 method, the highest scores can be obtained with different methods in the Turkish data set.

5. Conclusion and Future Works

In this paper, we have comprehensively analysed two datasets consisting of Turkish and English abstracts extracted from Turkish medical journals. A comprehensive study on classification of two counterparts abstracts was showed by using three classifiers. Three different globalisation techniques and four local feature selection methods

were used in performance analysis. Also, three pattern classifiers were used in classification stage. According to experimental studies, classification of English dataset containing medical abstracts is more successful than their counterparts in Turkish dataset. MNB classifier has gained more performance than SVM and DT classifiers on both data sets. The focal point of MNB classifier is the supposition of independence between terms. Also, MAX globalisation technique is the best method according to classification performance for most cases in both datasets. While CHI2 method is more successful than other methods, MI method is the worst method for most cases in both datasets. As a future work, a novel globalisation technique may be developed for medical domain. Also, the effect of different feature representation methods may be investigated in both languages. The analyzes made in the study were applied for two different data sets. However, more feature extraction methods and classification can be applied to different data. The article can be extended with the latest globalization techniques.

Table 3 Micro and Micro -F scores (%) obtained on English dataset with MNB

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	87.70	86.46	83.42	40.23	47.13	47.95	82.45	82.51	80.15	80.65	73.02	74.01
250	92.40	90.26	88.85	42.23	50.46	58.06	87.47	86.68	84.96	86.17	81.65	81.47
500	93.01	92.75	91.10	45.36	58.23	68.60	92.04	87.98	88.69	89.61	87.53	86.35
1000	94.89	94.70	93.81	48.46	63.77	77.56	94.40	89.94	90.15	91.88	90.26	89.67
2000	**95.48**	94.99	94.55	61.87	72.09	**85.83**	**95.14**	92.14	90.52	**93.16**	92.34	91.57
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	74.67	72.37	57.64	04.62	16.72	18.89	68.87	64.74	54.56	64.84	49.88	50.92
250	82.26	78.52	73.79	11.80	20.82	32.07	76.59	71.85	62.44	71.91	64.55	61.64
500	83.34	82.90	78.47	20.35	31.86	44.11	81.53	74.17	71.78	77.55	74.12	69.90
1000	87.71	87.20	84.04	26.88	40.86	54.56	86.73	77.08	76.38	81.45	78.43	76.70
2000	**88.25**	87.63	86.47	48.21	53.37	**66.68**	**87.42**	80.93	78.13	**84.15**	82.09	80.57

Table 4 Micro and Micro-F scores (%) obtained on English dataset with SVM

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	87.08	84.43	82.51	40.45	48.05	47.34	80.90	80.21	79.06	77.49	70.18	71.73
250	87.98	86.91	85.25	43.32	52.12	57.97	83.72	84.14	81.90	83.12	78.74	77.49
500	88.20	88.85	88.75	45.46	55.89	62.62	86.86	86.35	86.06	88.97	87.36	86.74
1000	91.73	92.19	91.62	48.15	58.41	68.68	87.59	88.14	88.20	91.52	91.10	91.52
2000	92.70	93.11	**93.26**	59.47	66.04	**79.45**	89.35	**91.21**	90.84	92.40	92.45	**92.60**
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	72.34	67.58	58.05	05.73	16.87	15.09	68.68	60.72	51.00	58.81	43.97	46.50
250	74.20	73.40	67.54	15.05	24.52	28.43	69.01	68.71	58.83	67.13	60.09	54.41
500	75.74	77.33	73.70	21.24	28.39	37.44	74.14	71.71	66.63	76.75	74.33	69.31
1000	80.87	81.93	80.47	29.06	34.16	42.38	72.41	74.30	73.65	80.55	79.41	79.57
2000	82.52	82.90	**84.64**	46.81	47.03	**56.94**	77.31	**79.73**	78.34	**82.35**	82.17	82.05

Table 5 Micro and Macro-F scores (%) obtained on English dataset with DT

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	**86.46**	84.02	81.90	40.11	45.46	48.87	79.25	80.21	77.02	76.76	65.88	69.21
250	85.02	83.60	85.13	41.23	45.78	55.88	82.39	80.84	81.47	82.02	77.69	73.80
500	85.31	85.71	85.13	41.57	47.75	62.29	85.71	81.90	81.40	83.78	82.57	80.91
1000	85.13	85.88	86.06	49.27	49.67	68.83	**85.77**	82.39	82.63	84.49	83.90	84.02
2000	85.77	85.60	85.37	57.34	59.20	**77.09**	85.19	85.19	84.31	**84.78**	84.49	84.61
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	**71.79**	67.91	57.78	04.01	11.91	15.55	66.87	62.84	49.02	59.43	39.49	42.08

250	68.40	66.20	68.00	08.12	12.43	23.44	67.34	62.46	56.26	64.74	58.50	45.93
500	69.93	70.13	68.87	09.23	20.08	33.77	**72.92**	64.91	57.48	67.06	64.66	60.19
1000	69.43	71.16	71.57	31.35	24.14	41.60	71.49	65.28	64.25	68.46	67.32	67.45
2000	70.88	70.78	70.35	44.88	34.17	**52.22**	70.26	69.45	67.07	**68.81**	67.87	68.41

Table 6 Micro and Macro-F scores (%) obtained on Turkish dataset with MNB

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	83.66	80.72	74.78	41.12	40.56	40.45	73.59	72.66	68.68	73.52	60.07	59.73
250	87.64	86.57	84.25	42.23	49.16	46.72	83.42	79.06	76.42	80.97	73.45	73.09
500	89.61	89.45	86.06	44.40	54.87	54.78	87.98	82.94	82.57	85.89	82.45	81.84
1000	91.41	91.67	89.99	48.26	62.62	54.22	90.09	87.36	85.89	89.29	88.14	86.57
2000	**92.75**	92.34	91.67	56.52	**79.25**	55.61	**91.42**	89.94	86.63	**92.08**	90.68	90.94
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	66.25	58.89	42.68	08.11	10.26	08.60	58.79	45.20	38.05	47.57	29.19	29.34
250	74.29	71.89	61.56	10.71	22.74	15.069	71.37	57.55	46.23	63.83	49.14	44.87
500	77.23	76.88	64.90	17.24	26.88	21.84	76.07	65.55	58.91	70.38	65.65	60.59
1000	80.97	81.23	74.61	24.88	35.79	22.14	77.43	71.17	62.42	76.49	74.92	70.30
2000	**82.82**	82.38	78.87	38.24	**55.80**	21.45	**79.87**	76.71	64.73	**80.70**	77.98	78.28

Table 7 Micro-F scores (%) obtained on Turkish dataset with SVM

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	81.84	79.25	73.24	40.90	41.12	46.40	74.29	72.02	69.74	71.58	59.12	60.68
250	83.54	83.30	80.47	42.01	49.47	52.51	80.15	74.22	76.76	74.01	66.98	67.21
500	83.30	82.76	81.40	43.97	51.74	55.15	80.84	77.42	78.02	79.19	76.28	75.81
1000	83.36	84.14	85.08	45.25	60.33	56.98	83.48	80.78	80.47	84.25	82.57	83.00
2000	84.78	**86.23**	85.83	52.60	**71.22**	29.33	81.03	**84.67**	83.42	86.69	87.03	**87.31**
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	62.91	57.94	42.99	08.26	10.27	09.30	54.75	46.29	37.40	45.18	27.94	30.00
250	64.83	65.33	58.20	11.36	20.12	18.07	62.02	49.55	51.62	54.20	42.60	42.00
500	65.30	64.30	59.19	16.39	23.27	24.34	62.58	56.64	52.08	60.90	57.23	53.52
1000	65.25	67.53	66.32	25.06	32.31	28.34	64.03	61.10	56.40	68.09	65.39	64.41
2000	66.41	68.18	**68.76**	35.26	46.37	**59.55**	60.37	**68.31**	62.41	71.10	71.52	**72.11**

Table 8 Micro-F scores (%) obtained on Turkish dataset with DT

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	79.64	76.76	73.59	39.89	40.56	44.51	70.18	68.83	69.51	66.67	52.80	51.15
250	**83.00**	81.34	78.61	40.11	45.46	52.60	77.82	73.59	75.95	76.08	63.29	64.59
500	81.40	79.96	80.08	42.23	46.51	53.75	80.59	74.22	76.15	77.89	74.50	72.23
1000	79.96	81.59	81.40	44.83	55.61	57.07	**83.66**	78.41	75.67	79.96	78.15	79.06
2000	82.27	81.09	81.53	47.34	**68.37**	59.73	81.96	79.57	77.02	**80.91**	79.56	79.96
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	59.16	54.14	44.47	04.26	08.22	07.13	48.02	44.27	38.21	40.68	25.10	25.13
250	**64.43**	60.76	53.38	05.37	14.85	16.96	59.73	49.09	49.50	55.67	37.28	36.19
500	62.20	59.30	58.12	11.26	17.17	20.79	62.52	50.59	50.52	55.14	52.20	46.45
1000	59.87	61.24	60.70	18.23	25.79	22.81	**64.75**	57.06	51.41	60.03	57.37	58.16
2000	62.24	60.59	61.21	28.93	**39.82**	24.26	61.33	57.51	53.09	**60.92**	60.23	59.87

References

[1] C. C. Aggarwal and C. Zhai, Mining text data. Springer Science & Business Media, 2012.

[2] G. V. Cormack, "Email spam filtering: A systematic review," 2008.

[3] S. J. Delany, M. Buckley, and D. Greene, "SMS spam filtering: Methods and data," Expert Systems with Applications, vol. 39, no. 10, pp. 9899-9908, 2012.

[4] M. Coulthard, "Author identification, idiolect, and linguistic uniqueness," Applied linguistics, vol. 25, no. 4, pp. 431-447, 2004.

[5] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye, "Author identification on the large scale," in Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA), 2005.

[6] S. M. Zu Eissen and B. Stein, "Genre classification of web pages," in Annual Conference on Artificial Intelligence, 2004, pp. 256-269: Springer.

[7] B. Choi and X. Peng, "Dynamic and hierarchical classification of web pages," Online Information Review, 2004.

[8] N. Isaacson, "The “fetus-infant”: Changing classifications of In Utero development in medical texts," in Sociological Forum, 1996, vol. 11, no. 3, pp. 457-480: Springer.

[9] B. Parlak and A. K. Uysal, "On classification of abstracts obtained from medical journals," Journal of Information Science, vol. 46, no. 5, pp. 648-663, 2020.

[10] N. L. Medicine. (2020, 2020-09-30). Available: https://pubmed.ncbi.nlm.nih.gov/

[11] C. A. Gonçalves, C. T. Gonçalves, R. Camacho, and E. C. Oliveira, "The Impact of Pre-processing on the Classification of MEDLINE Documents," in PRIS, 2010, pp. 53-61.

[12] M. Yetisgen-Yildiz and W. Pratt, "The effect of feature representation on MEDLINE document classification," in AMIA annual symposium proceedings, 2005, vol. 2005, p. 849: American Medical Informatics Association.

[13] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," Journal of biomedical informatics, vol. 53, pp. 196-207, 2015.

[14] J. G. Adeva, J. P. Atxa, M. U. Carrillo, and E. A. Zengotitabengoa, "Automatic text classification to support systematic reviews in medicine," Expert Systems with Applications, vol. 41, no. 4, pp. 1498-1508, 2014.

[15] N. Sánchez-Marono, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection–a comparative study," in International Conference on Intelligent Data Engineering and Automated Learning, 2007, pp. 178-187: Springer.

[16] L. Talavera, "An evaluation of filter and wrapper methods for feature selection in categorical clustering," in International Symposium on Intelligent Data Analysis, 2005, pp. 440-451: Springer.

[17] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded methods," in Feature extraction: Springer, 2006, pp. 137-165.

[18] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," Knowledge-Based Systems, vol. 36, pp. 226-235, 2012.

[19] V. Srividhya and R. Anitha, "Evaluating preprocessing techniques in text categorization," International journal of computer science and application, vol. 47, no. 11, pp. 49-51, 2010.

[20] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," ACM Sigkdd Explorations Newsletter, vol. 6, no. 1, pp. 80-89, 2004.

[21] H. Taira and M. Haruno, "Feature selection in SVM text categorization," in AAAI/IAAI, 1999, pp. 480-486.

[22] S. Günal, "Hybrid feature selection for text classification," Turkish Journal of Electrical Engineering and Computer Science, vol. 20, no. Sup. 2, pp. 1296-1311, 2012.

[23] G. BİRİCİK, B. Diri, and A. C. SÖNMEZ, "Abstract feature extraction for text classification," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 20, no. Sup. 1, pp. 1137-1159, 2012.

[24] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, "Text categorization: past and present," Artificial Intelligence Review, vol. 54, no. 4, pp. 3007-3054, 2021.

[25] M. Belazzoug, M. Touahria, F. Nouioua, and M. Brahimi, "An improved sine cosine algorithm to select features for text categorization," Journal of King Saud

University-Computer and Information Sciences, vol. 32, no. 4, pp. 454-464, 2020.

[26] H. Chantar, M. Mafarja, H. Alsawalqah, A. A. Heidari, I. Aljarah, and H. Faris, "Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification," Neural Computing and Applications, vol. 32, no. 16, pp. 12201-12220, 2020.

[27] B. M. Zahran and G. Kanaan, "Text feature selection using particle swarm optimization algorithm 1," 2009.

[28] F. Dong and Y. Zhang, "Automatic features for essay scoring–an empirical study," in Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 1072-1077.

[29] V. Dasondi, M. Pathak, and N. P. Singh, "An implementation of graph based text classification technique for social media," in 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016, pp. 1-7: IEEE.

[30] F. D. Malliaros and K. Skianis, "Graph-based term weighting for text categorization," in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015, pp. 1473-1479.

[31] L. Kumari, "Improved Graph Based K-NN Text Classification," Int J Eng Res Appl, vol. 3, pp. 928-931, 2013.

[32] C. Jiang, F. Coenen, R. Sanderson, and M. Zito, "Text classification using graph mining-based feature extraction," in Research and Development in Intelligent Systems XXVI: Springer, 2010, pp. 21-34.

[33] R. Liu, J. Zhou, and M. Liu, "Graph-based semi-supervised learning algorithm for web page classification," in Sixth International Conference on Intelligent Systems Design and Applications, 2006, vol. 2, pp. 856-860: IEEE.

[34] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in Icml, 1997, vol. 97, no. 412-420, p. 35: Nashville, TN, USA.

[35] R. A. Calvo and H. A. Ceccatto, "Intelligent document classification," Intelligent Data Analysis, vol. 4, no. 5, pp. 411-420, 2000.

[36] T. DOĞAN and A. K. UYSAL, "The Effects of Globalisation Functions on Feature Weighting for Text Classification," in 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018, pp. 1-6: IEEE.

[37] B. Parlak and A. K. Uysal, "A novel filter feature selection method for text classification: Extensive Feature Selector," Journal of Information Science, p. 0165551521991037, 2021.

[38] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, Introduction to pattern recognition: a matlab approach. Academic Press, 2010.

[39] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in European conference on machine learning, 1998, pp. 137-142: Springer.

[40] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," Applied Soft Computing, vol. 86, p. 105836, 2020.

[41] A. K. Uysal, "An improved global feature selection scheme for text classification," Expert systems with Applications, vol. 43, pp. 82-92, 2016.

[42] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," Machine learning, vol. 39, no. 2, pp. 103-134, 2000.