Effects of Preprocessing on Text Classification in Balanced and Imbalanced Datasets

Mehmet F. Karaca*
Department of Management Information Systems, Erbaa Faculty of Social Sciences and Humanities, Tokat Gaziosmanpaşa University, Post Code: 60500, Erbaa, Tokat, Turkey
[e-mail: mehmetfatih.karaca@gop.edu.tr]
*Corresponding author: Mehmet F. Karaca

Received May 30, 2022; revised July 28, 2023; revised August 30, 2023; accepted February 26, 2024; published March 31, 2024

Abstract

In this study, preprocessings with all combinations were examined in terms of the effects on decreasing word number, shortening the duration of the process and the classification success in balanced and imbalanced datasets which were unbalanced in different ratios. The decreases in the word number and the processing time provided by preprocessings were interrelated. It was seen that more successful classifications were made with Turkish datasets and English datasets were affected more from the situation of whether the dataset is balanced or not. It was found out that the incorrect classifications, which are in the classes having few documents in highly imbalanced datasets, were made by assigning to the class close to the related class in terms of topic in Turkish datasets and to the class which have many documents in English datasets. In terms of average scores, the highest classification was obtained in Turkish datasets as follows: with not applying lowercase, applying stemming and removing stop words, and in English datasets as follows: with applying lowercase and stemming, removing stop words. Applying stemming was the most important preprocessing method which increases the success in Turkish datasets, whereas removing stop words in English datasets. The maximum scores revealed that feature selection, feature size and classifier are more effective than preprocessing in classification success. It was concluded that preprocessing is necessary for text classification because it shortens the processing time and can achieve high classification success, a preprocessing method does not have the same effect in all languages, and different preprocessing methods are more successful for different languages.

Keywords: Natural Language Processing, Pattern Recognition, Preprocessing, Text Classification, Text Mining.

1. Introduction

Data including various types of contents in digital environment reached enormous dimensions today. Textual data is one of the types of these data. Text classification is assigning documents to predefined classes. Web page classification [1, 2], sentiment classification [3-5], customer complaints classification [6, 7], spam detection [8, 9], tweet classification [10, 11] and other classifications [6, 12-18] are samples of text classification in digital environment.

Text classification contains preprocessing, feature selection, feature size, term weighting, obtaining document vectors and classification processes, and these processes also include sub-processes. Researchers consider the determination of the factors which affect the classification recently. How more successful classifications will be made in a shorter time can be revealed as a result of determining key factors and improving these factors.

Preprocessing is one of the basic steps of text classification [19]. In preprocessing process, one or more following steps can be included: data cleaning, tokenization, lowercase conversion, stemming application and/or stop words removal [1, 5, 6, 8-18, 20-28]. With feature selection, determination of the terms, which are going to be used in the classification out of the words which compose the document, with document frequency [19, 27], Chi-square [19, 23, 24, 27, 28], odds ratio [19, 27], mutual information [19, 23], information gain [6, 7, 9, 19, 23, 27] or with other methods [5, 15, 19, 23, 29] is provided. With feature size, feature set is generally constructed by extracting the terms in equal and specific numbers from each class, and the feature set is matched with the document vectors. As a result of this match, document vectors are obtained by applying term weightings as binary [20], term frequency [20, 22], term frequency-inverse document frequency [4, 5, 7, 17-22, 24, 25] or other weighting methods [20, 24, 25]. Finally, classification is made with kNN (k-Nearest Neighbors) [1, 5, 6, 9, 17, 19, 22, 24, 25], Naive Bayes [5, 7, 8, 10, 12, 13, 17-20, 22, 27], SVM (Support Vector Machines) [1, 5, 7-14, 17-20, 22-25, 27, 28] or other classifiers [1-3, 5-10, 12, 14-18, 20, 22, 23, 25, 26, 29]. Techniques which are applied in order to reveal the effects of the text classification processes are tested with datasets which are considered as standard datasets such as 20Newsgroups [1, 14, 19-21, 24, 27] or Reuters-21578 [19, 20, 24, 28]; therefore, it is aimed to provide an opportunity in order to make a general evaluation.

Preprocessing which is also the topic of this study varies according to the source or language of the document, or the analysis desired to be done [1, 6, 8-13, 20-28]. There are a few studies in which the preprocessing, which is applied as intermediary process in many studies, is taken into consideration from various perspectives. In one of these studies [28], classifications were made with various preprocessing methods in Turkish and English datasets and the following results were obtained: the highest classification was obtained with alphabetic tokenization as 0.971 in Turkish dataset when stop words and stemming are not applied, and lowercase conversion is applied, and the highest classification was also obtained with alphabetic tokenization as 0.989 in English dataset when stop words is not applied, and lowercase conversion and stemming are applied. Binary classification was done with the documents included in 2 Turkish datasets in the study [26] carried out to reveal the effects of preprocessings on text classification success. The following series of processes were determined as preprocessing: removing punctuation marks, lowercase conversion, deleting number, preposition, conjunction, money, weight and length expressions, correcting spelling mistakes, changing link information into url and username into usr. It was stated that applying preprocessing has a positive effect and preprocessing has a significant effect in the dataset which has more document number. Turkish document classification was done in 6-class balanced dataset in the study [22] which included the stages of preprocessing. Data correction

(correcting mistakes/hyphenated words), stop words removal and lemmatization were applied as preprocessing. It was observed that the classification success increases less when the data correction is applied, besides, it increases relatively more when the stop words are removed, but the classification success decreases when the lemmatization is applied. The highest classification was obtained as 0.918 with SVM (linear kernel) when the stop words are removed and the lemmatization is not applied. In the study [24] in which the preprocessing was applied as an intermediary process, multiclass classification was done with 1 balanced and 2 imbalanced English datasets. Lowercase conversion, alphabetic tokenization, stop words removal and stemming preprocessings were applied. It was seen that SVM (approximately 0.965 with linear kernel) classify the documents generally better than kNN. In the text classification study [23], the classification was done by using 6-class imbalanced datasets including Arabic texts. The following preprocessing stages were applied: digits, punctuation marks, number, non-Arabic characters, stop words and non-useful words were deleted, some letters were replaced. The highest classification was obtained as 0.905 when SVM classifier and improved Chi-square feature selection are applied together. In other classification study [21], 20-class imbalanced English dataset was used. Firstly, tokenization, then stop words removal, and finally stemming preprocessings were applied. The highest success was obtained in the class having maximum documents as 0.908.

1.1 Research Questions and Organization

When the relevant studies are considered, it was seen that generally only one factor which affect the classification is taken into consideration, the interdependent effects of these factors were not studied in detail, and also the effects of preprocessing methods on the classification success in balanced and imbalanced datasets are not examined comprehensively.

In the proposed work, it is specifically aimed to reveal the effects of preprocessing on text classification in balanced and imbalanced datasets. Classifications were carried out on Turkish and English datasets (5 datasets for each language) having different imbalance levels with all combinations of lowercase conversion, stemming application and stop words removal which are the preprocessing methods. By this way, it is aimed to find out answers for the following issues and to discuss the issue in detail:

- What is the effect of preprocessing on reducing the number of discrete and total words?
- Which preprocessing method(s) is/are more effective in text classification?
- In which way (positive/negative) and to what extent do preprocessing methods have an effect on the success of text classification in balanced/imbalanced datasets?
- How does the imbalance of datasets affect text classification success?
- What is the classification tendency in classes with few documents?
- Does the applied preprocessings classification success differ from language to language?
- What are the effects of preprocessings in terms of processing time?
- How are feature selection, feature size and classification algorithms, which are other factors affecting classification, affected by preprocessings?
- Are these factors more effective in classification than preprocessings?
- Is it necessary to apply preprocessing when evaluated in terms of processing time and classification success?

The rest of this paper is organized as follows: In Section 2, the methodology is described. In Section 3, experimental results are provided and discussed in comparison with the literature. Finally, in Section 4 the study is concluded.

2. Methodology

The effects of preprocessing on text classification in balanced and imbalanced datasets were examined in this study. 8 sub-datasets were obtained as a result of all possible preprocessing combinations from each dataset, and 3 feature selections, 3 feature sizes, 3 kNN (as k=3, 5, 7), Multinomial Naive Bayes (MNB) and SVM classifiers were applied to these sub-datasets. It is aimed to conduct a comprehensive study with a total of 7920 classifications obtained from the following classification numbers: 99 classifications with a sub-dataset, 792 classifications with a dataset and 3960 classifications in a language.

As it is seen in **Fig. 1**, documents belong to each dataset are presented as input to the system. Documents are preprocessed, then the terms are determined depending on feature selection and feature size, and document vectors are generated by matching these terms with the documents. Finally, classification is obtained by applying classification algorithms to vectors.

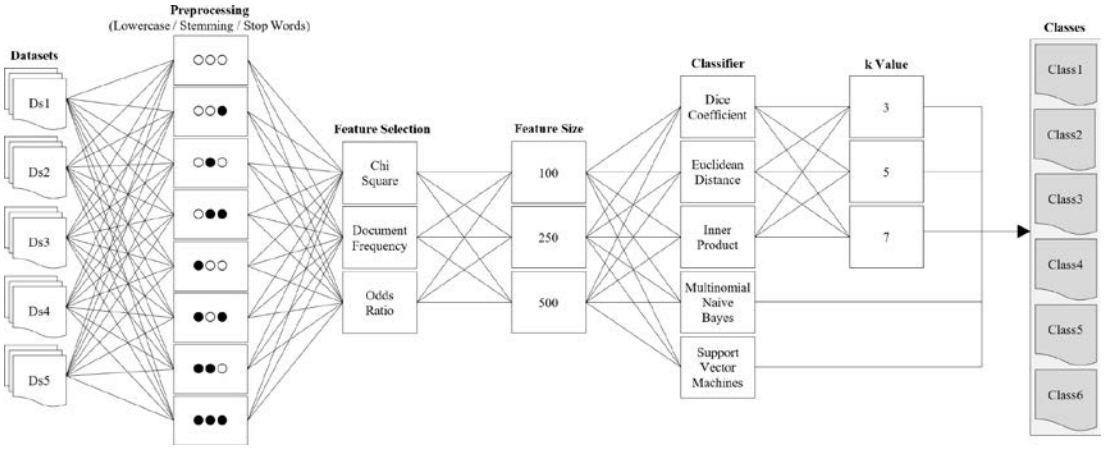


Fig. 1. Structure of the study

2.1 Datasets

Information related to 6-class datasets consisting of Turkish and English documents used in the study and their features are seen in **Table 1** and **Table 2**. Turkish datasets consisting of the articles of newspaper columnists were labelled as T-Ds{1-5} and English datasets consisting of the documents of 20Newsgroups were labelled as E-Ds{1-5}. 20Newsgroups dataset is a 20-class dataset including approximately 900 documents in each class and is frequently used on text classification studies. Classes which are close to T-Ds dataset in terms of subject matter were tried to be selected among 20Newsgroups dataset.

Since T-Ds2 is the dataset which includes the maximum number of documents, T-Ds1, T-Ds3, T-Ds4 and T-Ds5 datasets were generated from this dataset for Turkish. The only balanced dataset is T-Ds1. The imbalanced ratio is gradually increasing from T-Ds2 to T-Ds5. For instance, the document number ratios of classes to total document number are in the range of 12.50% - 18.75% in T-Ds2, on the other hand, 1.52% - 45.69% in T-Ds5.

Since there is not enough document in terms of number in classes of 20Newsgroups dataset, E-Ds datasets and T-Ds datasets were conformed in terms of proportional but not numerical data. For instance, the dataset which is balanced on T-Ds is also balanced on E-Ds, the imbalanced ratio of T-Ds datasets and same E-Ds datasets are equal.

Table 1. Datasets and sample numbers generated from Turkish dataset

	T-Ds1		T-Ds2		T-Ds3		T-Ds4		T-Ds5	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**
Politics	900	300	900	300	750	250	450	150	75	25
Education	900	300	1200	400	1050	350	900	300	525	175
Economy	900	300	1350	450	1350	450	1350	450	1350	450
Health	900	300	1350	450	1200	400	1125	375	720	240
Sports	900	300	1200	400	900	300	675	225	240	80
Life	900	300	1200	400	600	200	225	75	45	15
Total	5400	1800	7200	2400	5850	1950	4725	1575	2955	985

Table 2. Datasets and sample numbers generated from English dataset

	E-Ds1		E-Ds2		E-Ds3		E-Ds4		E-Ds5	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**
talk.politics.mideast	450	150	450	150	375	125	225	75	38	12
sci.crypt	450	150	600	200	525	175	450	150	262	88
sci.space	450	150	675	225	675	225	675	225	675	225
sci.med	450	150	675	225	600	200	562	188	360	120
rec.sport.hockey	450	150	600	200	450	150	338	112	120	40
soc.religion.christian	450	150	600	200	300	100	112	38	22	8
Total	2700	900	3600	1200	2925	975	2362	788	1477	493

2.2 Preprocessing

Series of processes must be executed for documents to carry out text classification. It can be said that preprocessing is the first stage of this process. Preprocessing is the procedure of processing textual data depending on various criteria and obtaining pure text. In this process, there are operations such as data cleaning, tokenization, lowercase conversion, stemming application and stop words removal, etc.

Data cleaning is the process of removing unnecessary ones for the study from textual data. Data cleaning displays differences according to data. For instance, web statements must be removed if they are unnecessary for the study which is going to be conducted with web documents or punctuation marks and unnecessary spaces which do not make contribution to the study must be cleaned. ***Tokenization*** is splitting text into smaller units which are called as tokens (words). After tokenization, texts are processed in the word level. ***Lowercase*** is the conversion of a word into a lowercase. ***Stemming*** is obtaining the stem of the word by removing affixes. Different stemming algorithms are applied for each language. As in many studies, for Turkish texts Zemberek stemmer [30] and for English texts Porter stemmer [31] was used in this study. ***Stop words*** are the words which are frequently included within texts; therefore, they are considered as the words which do not contribute to classification. It is stated that stop words have negative effect on classification success and the success will be increased by removing these from texts [22].

Preprocessing processes applied to a dataset within the scope of the study are below:

```
for each doc in dataset
  pureDoc=DataCleaning(doc)
  words[ ]=Tokenization(pureDoc,‘ ’)
  for each word in words
    if Lowercase
      word=LowercaseConversion(word)
    if Stemming
      word=StemmingApplication(word)
    if StopWordsRemoval
      if word in stopWordsList{...}
        continue for
    finalDoc=finalDoc & ‘ ’ & word
```

In this study, all variations of preprocessing were tested with all datasets. In **Table 3**, the following marks are used respectively as: ✓ applied and × not applied, to show how these variations were obtained. Moreover, LC denotes lowercase conversion, ST stemming application and SW stop words removal. For instance, it is stated that ○●○$_{LC\times ST\checkmark SW\times}$ marking denotes that: LC not applied, ST applied and SW not applied; ●○●$_{LC\checkmark ST\times SW\checkmark}$ marking denotes that: LC applied, ST not applied and SW applied.

Table 3. Applied preprocessing methods

	LC	ST	SW
○○○$_{LC\times ST\times SW\times}$	×	×	×
○○●$_{LC\times ST\times SW\checkmark}$	×	×	✓
○●○$_{LC\times ST\checkmark SW\times}$	×	✓	×
○●●$_{LC\times ST\checkmark SW\checkmark}$	×	✓	✓
●○○$_{LC\checkmark ST\times SW\times}$	✓	×	×
●○●$_{LC\checkmark ST\times SW\checkmark}$	✓	×	✓
●●○$_{LC\checkmark ST\checkmark SW\times}$	✓	✓	×
●●●$_{LC\checkmark ST\checkmark SW\checkmark}$	✓	✓	✓

Analyses of Turkish and English sentence samples which have the same meaning obtained depending on the preprocessings applied within the framework of the study are given in **Table 4**. Only punctuation marks were removed from the text in ○○○$_{LC\times ST\times SW\times}$. Whereas the number of discrete and total word is 10 in Turkish, 14 in English. After lowercase conversion (for instance, ●○○$_{LC\checkmark ST\times SW\times}$), the same English words “*Health*” and “*health*” were processed as “*health*”. When stemming was applied (for instance, ○●○$_{LC\times ST\checkmark SW\times}$), Zemberek stemmer present the stem that it found into lowercase, on the other hand, Porter stemmer presented it as in its original form. For instance, while the words “*Sağlık*” and “*sağlığına*” in Turkish sentences were presented as only one stem “*sağlık*” by Zemberek stemmer, the words “*Health*” and “*health*” in English sentences were presented as two different stems “*Health*” and “*health*” by Porter stemmer. Moreover, Zemberek stemmer could not determine the roots of proper names converted into lower case. For instance, the word “*mehmet*,” which is a proper name in ●○○$_{LC\checkmark ST\times SW\times}$, could not be determined by Zemberek stemmer in terms of roots, therefore, it was not included in ●●○$_{LC\checkmark ST\checkmark SW\times}$ which is the root of the word. With the removal of stop words more decrease was obtained in the total word number in English compared to Turkish (for instance, ○○●$_{LC\times ST\times SW\checkmark}$).

Table 4. Turkish and English sentences analyses of preprocessing

	Turkish Sentences	**English Sentences**
	Sağlık her şeyden önemli. Fakat Mehmet sağlığına hiç dikkat etmiyor.	*Health is more important than everything. But Mehmet never takes care of his health.*
○○○LC✖ST✖SW✖	Sağlık her şeyden önemli Fakat Mehmet sağlığına hiç dikkat etmiyor	Health is more important than everything But Mehmet never takes care of his health
○○●LC✖ST✖SW✓	Sağlık şeyden önemli Fakat Mehmet sağlığına dikkat etmiyor	Health important But Mehmet takes care health
○●○LC✖ST✓SW✖	sağlık her şey önem fakat mehmet sağlık hiç dikkat et	Health is more import than everyth But Mehmet never take care of hi health
○●●LC✖ST✓SW✓	sağlık önem mehmet sağlık dikkat et	Health import Mehmet take care health
●○○LC✓ST✖SW✖	sağlık her şeyden önemli fakat mehmet sağlığına hiç dikkat etmiyor	health is more important than everything but mehmet never takes care of his health
●○●LC✓ST✖SW✓	sağlık şeyden önemli mehmet sağlığına dikkat etmiyor	health important mehmet takes care health
●●○LC✓ST✓SW✖	sağlık her şey önem fakat sağlık hiç dikkat et	health is more import than everyth but mehmet never take care of hi health
●●●LC✓ST✓SW✓	sağlık önem sağlık dikkat et	health import mehmet take care health

2.3 Feature Selection and Feature Size

Feature selection is the process of selecting by determining distinctive terms which represent the class best from training documents. CHI (Chi-Square), DF (Document Frequency) and OR (Odds Ratio) feature selection methods were used in this study. Feature set was generated by selecting 100, 250 and 500 terms from each class and matched with documents, then document vectors are obtained with binary weighting. If the term is found in the document, this case is represented with 1, if not, with 0 in vectors generated by binary weighting. Preprocessing and feature selection decrease the dimension of the vector; therefore, process duration is shortened.

2.4 Classification

Classification is made with kNN, MNB and SVM in this study. Dice (Dice Coefficient) (1), Euclid (Euclidean Distance) (2) and Inner (Inner Product) (3) methods which are used to measure the similarities between the testing document vector and training document vectors in kNN were tested with the values of 3, 5 and 7 *k*. Whether any term in Euclid exists or does not exist both in testing and training vectors, it has the same value. While it is an important issue for the term to exist in both two vectors at the same time in Inner; besides, vector lengths are also taken into consideration in Dice. *X* and *Y* represent vectors, *i* indicates term, *n* stands for total term number and *d* denotes similarity measurement of vectors.

$$d(X,Y)=\frac{2\times\sum_{i=1}^{n}(X_i\times Y_i)}{\sum_{i=1}^{n}(X_i)^2+\sum_{i=1}^{n}(Y_i)^2} \tag{1}$$

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(X_i-Y_i)^2} \tag{2}$$

$$d(X,Y)=\sum_{i=1}^{n}(X_i\times Y_i) \tag{3}$$

In MNB, which is a probabilistic method, primarily, probability value is calculated for each class, after that, class which has the highest probability value is assigned as the class of test document [7, 8, 22]. Moreover, linear kernel function is preferred in SVM as a result of obtaining more successful results [8, 11, 22].

2.5 Success Measure

Micro-F1 and Macro-F1 scores are generally used to measure performance in text classification studies. In the calculation of Macro-F1 score, averages of F values obtained for classes are calculated and this causes to consider all classes at the same ratio. That Macro-F1 considers class which has few documents with class that has many documents at the same level in imbalanced datasets can prevent performing correct evaluation in the studies especially carried out with highly imbalanced datasets. Since the datasets used in our study are mostly imbalanced, *Micro-F1* (6) score is preferred in the measurement of classification performance. Firstly, *tp* (true positive), *fp* (false positive) and *fn* (false negative) values were obtained in order to calculate *p* (precision) (4) and *r* (recall) (5) values. Then, Micro-F1 (6) value was calculated with *p* and *r* values. Besides, *tp* denotes document number assigned to the correct class, *fp* denotes document number assigned which does not even belong to the class, *fn* denotes document number which was not assigned which even belongs to the class.

$$p = \frac{tp}{tp + fp} \tag{4}$$

$$r = \frac{tp}{tp + fn} \tag{5}$$

$$Micro - F1 = \frac{2 \times p \times r}{p + r} \tag{6}$$

3. Experimental Results

In this study, lowercase conversion, application of stemming and removal of stop words which are the procedures of preprocessing processes, with all their combinations are tested in 1 balanced dataset and 4 imbalanced datasets having different imbalance levels. 8 sub-datasets including the preprocessing combinations of each dataset were applied with their sub-dimensions of feature selection, feature size and classifier having different features; therefore, it is aimed to reveal the effects of preprocessing on text classification.

At first, number of the discrete and the total words obtained depending on preprocessing in datasets were evaluated. In order to reveal the effects of preprocessing on reducing the number of discrete and total words, the applied preprocessings were examined with ○○○$_{LC\times ST\times SW\times}$ comparatively, in which preprocessing was not applied. In addition, the comparison was made on the TDs-2 and Eds-2 datasets having maximum documents. The results revealed that, except for TDs-2 and Eds-2, other datasets vary almost (more or less) depending on the imbalance rate of the dataset.

The average decrease in the number of discrete words obtained as a result of preprocessing applied to T-Ds2 datasets in terms of % is as follows: ○○●$_{LC\times ST\times SW\checkmark}$ 0, ○●○$_{LC\times ST\checkmark SW\times}$ and ○●●$_{LC\times ST\checkmark SW\checkmark}$ 95, ●○○$_{LC\checkmark ST\times SW\times}$ and ●○●$_{LC\checkmark ST\times SW\checkmark}$17, ●●○$_{LC\checkmark ST\checkmark SW\times}$ and ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ 95, and the average decrease in the number of total words in T-Ds2 datasets in terms of % is as follows: ○○●$_{LC\times ST\times SW\checkmark}$ 24, ○●○$_{LC\times ST\checkmark SW\times}$ 5, ○●●$_{LC\times ST\checkmark SW\checkmark}$ 34, ●○○$_{LC\checkmark ST\times SW\times}$ 0, ●○●$_{LC\checkmark ST\times SW\checkmark}$ 25, ●●○$_{LC\checkmark ST\checkmark SW\times}$ 7 and ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ 37. The average decrease in the number of discrete words in E-Ds2 datasets in terms of % is as follows: ○○●$_{LC\times ST\times SW\checkmark}$ 0, ○●○$_{LC\times ST\checkmark SW\times}$ 25, ○●●$_{LC\times ST\checkmark SW\checkmark}$ 26, ●○○$_{LC\checkmark ST\times SW\times}$ 16, ●○●$_{LC\checkmark ST\times SW\checkmark}$ 17, ●●○$_{LC\checkmark ST\checkmark SW\times}$ 40, ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ 41, and the average decrease in the number of total words in E-Ds2 datasets in terms of % is as follows: ○○●$_{LC\times ST\times SW\checkmark}$ 50, ○●○$_{LC\times ST\checkmark SW\times}$ 5, ○●●$_{LC\times ST\checkmark SW\checkmark}$ 47, ●○○$_{LC\checkmark ST\times SW\times}$ 0, ●○●$_{LC\checkmark ST\times SW\checkmark}$ 50, ●●○$_{LC\checkmark ST\checkmark SW\times}$ 5 and ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ 47.

The differences were found out in T-Ds and E-Ds datasets in terms of the number of the discrete and the total words. It can be said that the language affects the results obtained. For instance, the number of the use of Turkish stop words are less than English stop words. Since

Turkish is an agglutinative language, morphemes are not used as discrete words in a sentence, they are used as agglutinated to the words. Turkish sentence "Bugün akşam evde kitabını okuyacağım." and English sentence "*I will* read *your* book *at* home tonight." have the same meaning. There is no stop word in this Turkish sentence, on the other hand, half of the words (italic ones) are stop words in English sentence. Therefore, the following decreases were provided in the number of total words with ○○●$_{LC\times ST\times SW\checkmark}$: 24% in Turkish T-Ds2 dataset and 50% in English E-Ds2 dataset. Besides, as a result of stemming, more decrease was found out in Turkish T-Ds datasets compared to English E-Ds datasets in terms of the number of the discrete words. Decrease in the number of the discrete words can be accepted as the indicator of accessing to the stems of the words better. The levels of decreases obtained as a result of applying ○●○$_{LC\times ST\checkmark SW\times}$ to documents in discrete word number were as follows: 95% with Zemberek stemmer and 25% with Porter stemmer. The reasons for this are that Turkish and English languages have different structures, Zemberek, the Turkish natural language processing library, and Porter for English try to access stems with different algorithms. Zemberek stemmer determined the stem of the Turkish words "*gitmek, gittim*" ("gitmek": 'go', and "gittim": Past simple of 'go' as 'went') as one stem "*git*"; on the other hand, Porter stemmer determines as separate stems "*go*" and "*went*". Moreover, Porter stemmer determined a stem for each meaningful or meaningless word included in E-Ds documents. For instance, it even assigned a stem for a meaningless word "*AAAGGGHHH*" included in a document. For these reasons, decrease ratio of the number of the discrete word as a result of stemming was higher in Zemberek stemmer compared to Porter stemmer. Duration of preprocessing processes were also examined and it was observed that the decrease in word number and processing time as a result of preprocessings display parallel results. For instance, the average decrease in processing duration in T-Ds2 dataset in terms of % is as follows: ○○●$_{LC\times ST\times SW\checkmark}$ 1, ○●○$_{LC\times ST\checkmark SW\times}$ 81, ○●●$_{LC\times ST\checkmark SW\checkmark}$ 82, ●○○$_{LC\checkmark ST\times SW\times}$ 4, ●○●$_{LC\checkmark ST\times SW\checkmark}$ 5, ●●○$_{LC\checkmark ST\checkmark SW\times}$ 89, ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ 92. The highest decrease in terms of duration was provided with stemming in T-Ds datasets in which the highest decrease was also observed in the number of total words, and stop words removal in E-Ds datasets.

Information related to maximum Micro-F1 scores obtained in the classifications made with datasets and feature selection, feature size and classifier (kNN with *k* values) used in these classifications are given in **Table 5** for T-Ds datasets and in **Table 6** for E-Ds datasets. The most successful preprocessing methods and number of occurrences in T-Ds datasets were respectively as follows: ○●○$_{LC\times ST\checkmark SW\times}$ 1, ●○○$_{LC\checkmark ST\times SW\times}$ 3, ●○●$_{LC\checkmark ST\times SW\checkmark}$ and ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ 1. The most successful preprocessing methods and number of occurrences in E-Ds datasets were respectively as follows: ○○●$_{LC\times ST\times SW\checkmark}$ and ○●○$_{LC\times ST\checkmark SW\times}$ 1, ●○●$_{LC\checkmark ST\times SW\checkmark}$ 2, ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ 3. The way and number of obtaining maximum Micro-F1 scores in T-Ds datasets are as follows: CHI 51, DF 16, OR 4, 100 1, 250 17, 500 53, Dice 69, Inner 1, *k*=3 14, *k*=5 34, *k*=7 22 times. The way and number of obtaining maximum Micro-F1 scores in E-Ds datasets are as follows: CHI 17, OR 29, 100 1, 250 13, 500 32, Dice 35, Inner 11, *k*=3 32, *k*=5 7, *k*=7 7 times. Euclid, MNB and SVM were not seen in maximum scores of T-Ds datasets. DF, Euclid, MNB and SVM were not seen in maximum scores of E-Ds datasets.

In summary, different results were obtained in T-Ds and E-Ds datasets and even among their own datasets in terms of maximum Micro-F1 scores. For instance, maximum Micro-F1 scores in T-Ds datasets were not affected from being balanced or imbalanced dataset or the preprocessing applied as much as E-Ds datasets. Differences in maximum scores obtained as a result of preprocessing in all datasets of T-Ds were rather at the low level. For instance, 0.985 which is the lowest maximum Micro-F1 score in T-Ds1 was obtained with ○○○$_{LC\times ST\times SW\times}$ and ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ preprocessing methods, 0.988 which is the highest maximum Micro-F1 score in

T-Ds1 was obtained with ○●○$_{LC\times ST\checkmark SW\times}$ and ●○●$_{LC\checkmark ST\times SW\checkmark}$ preprocessings. A case which is similar to the case in T-Ds1 was also occurred in maximum Micro-F1 scores in T-Ds{2-5} datasets.

It can be said that whether applying preprocessing or not has an effect in low level (less than 1%) in obtaining maximum Micro-F1 score in Turkish text classification studies besides whether the dataset is balanced or not. But it is not right to say that for E-Ds datasets. It is determined that whether the dataset is balanced or not and preprocessing applied have effect on the results of E-Ds datasets. For instance, the lowest and highest maximum Micro-F1 scores and differences in E-Ds datasets are respectively as follows: in E-Ds1 0.943 ●●○$_{LC\checkmark ST\checkmark SW\times}$ and 0.959 ○○●$_{LC\times ST\times SW\checkmark}$ (1.70%), in E-Ds2 0.945 ○○○$_{LC\times ST\times SW\times}$ and 0.964 ●○●$_{LC\checkmark ST\times SW\checkmark}$ (2.01%), in E-Ds3 0.942 ○○○$_{LC\times ST\times SW\times}$ and 0.957 ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ (1.59%), in E-Ds4 0.938 ○○○$_{LC\times ST\times SW\times}$ and 0.956 ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ (1.92%), in E-Ds5 0.923 ○○○$_{LC\times ST\times SW\times}$ and 0.939 ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ (1.73%).

The way of how the classification is made with the testing documents included in the classes having few documents was discussed, and whether the documents are assigned to the class close to the topic of the related class or to the class which includes more documents was also discussed in highly imbalanced T-Ds5 and E-Ds5 datasets. Examination is made on the most successful classification (0.992; ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$) in T-Ds5 dataset. *Life* and *politics* are the two classes which have few numbers of testing documents. When the test results related to these two classes are examined, it is found out that 55% of the documents in *life* class and 43% of the documents in *politics* class are classified as correct, and also it is found out that the general features of the classes assigned as incorrectly are the classes which are close to the related classes in terms of topic. Examination is also made on the most successful classification (0.939; ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$) in E-Ds dataset as is the case in T-Ds5. *soc.religion.christian* and *talk.politics.mideast* are the two classes which have few documents. According to the test results related to these two classes, 27% of the documents in *soc.religion.christian* class and 38% of the documents in *talk.politics.mideast* class are classified as correct. Unlike T-Ds5, it is found out that the incorrect classifications are generally in the direction of the classes which have the highest document number, and the classifications are made in the direction of classes which are dominant in terms of number.

Information related to average Micro-F1 scores obtained in the classifications made with preprocessing methods are given in **Table 7** for T-Ds datasets and **Table 8** for E-Ds datasets. It was seen that the maximum average Micro-F1 scores were generally obtained with ○●●$_{LC\times ST\checkmark SW\checkmark}$ (0.930) and the minimum average scores were generally obtained with ○○●$_{LC\times ST\times SW\checkmark}$ (0.871) in T-Ds datasets; the maximum scores were generally obtained with ●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ (0.785) and the minimum average Micro-F1 scores were generally obtained with ○○○$_{LC\times ST\times SW\times}$ (0.742) in E-Ds datasets. It was determined that the minimum average Micro-F1 score (0.884) is obtained with balanced dataset T-Ds1; imbalanced ratio increases with the increase in classification success in T-Ds datasets. It was determined that the minimum average Micro-F1 score (0.755) is obtained with balanced dataset E-Ds1; imbalanced ratio generally increases with the decrease in classification success in E-Ds datasets.

According to average Micro-F1 scores of preprocessings, it was seen that success increases in each preprocessing, where stemming is applied, when lowercase conversion is applied on its own and with stop words removal, on the other hand, it was seen that success decreases when stop words removal is applied on its own in T-Ds datasets. It was generally seen that the use of double preprocessings in E-Ds datasets increases the success compared to single uses (more successful classifications were obtained with the following structures respectively: ○●●$_{LC\times ST\checkmark SW\checkmark}$ compared to ○○●$_{LC\times ST\times SW\checkmark}$ and ○●○$_{LC\times ST\checkmark SW\times}$, ●○●$_{LC\checkmark ST\times SW\checkmark}$ compared to ○○●$_{LC\times ST\times SW\checkmark}$ and ●○○$_{LC\checkmark ST\times SW\times}$, ●●○$_{LC\checkmark ST\checkmark SW\times}$ compared to ○●○$_{LC\times ST\checkmark SW\times}$ and ●○○$_{LC\checkmark ST\times SW\times}$), and triplets are more successful than doubles (●●●$_{LC\checkmark ST\checkmark SW\checkmark}$ compared to ○●●$_{LC\times ST\checkmark SW\checkmark}$, ●○●$_{LC\checkmark ST\times SW\checkmark}$ and ●●○$_{LC\checkmark ST\checkmark SW\times}$).

Table 5. Maximum Micro-F1 scores of T-Ds and the methods utilized

	T-Ds1	**T-Ds2**	**T-Ds3**	**T-Ds4**	**T-Ds5**
○○○$_{LC*ST*SW*}$	0.985 CHI-500-Dice-3 CHI-500-Dice-5	0.987 CHI-500-MNB	0.988 CHI-500-Dice-5 CHI-500-Dice-7	0.992 CHI-500-Dice-5	0.984 CHI-500-Dice-5 OR-500-Dice-3 OR-500-Dice-5
○○●$_{LC*ST*SW✓}$	0.986 DF-500-Dice-5 DF-500-Dice-7	0.985 CHI-250-Dice-3 CHI-250-Dice-5	0.987 CHI-500-Dice-3	0.986 CHI-500-Dice-5	0.987 DF-500-Dice-3 DF-500-Dice-5
○●○$_{LC*ST✓SW*}$	**0.988** CHI-500-Dice-5	0.985 CHI-250-Dice-5	0.985 CHI-250-Dice-5 CHI-250-Dice-7 CHI-500-Dice-5 DF-500-Dice-7	0.990 CHI-100-Dice-5 CHI-250-Dice-5 CHI-250-Dice-7 CHI-500-Dice-5 CHI-500-Dice-7	0.988 CHI-250-Inner-5 CHI-250-Dice-5 OR-500-Dice-5
○●●$_{LC*ST✓SW✓}$	0.986 CHI-500-Dice-5 CHI-500-Dice-7	0.984 DF-500-Dice-7	0.986 DF-500-Dice-7	0.990 DF-500-Dice-7	0.991 CHI-250-Dice-3
●○○$_{LC✓ST*SW*}$	0.987 CHI-500-Dice-5 CHI-500-Dice-7	**0.988** CHI-500-Dice-5	**0.989** CHI-500-Dice-7	**0.993** CHI-500-Dice-7	0.985 OR-500-Dice-3
●○●$_{LC✓ST*SW✓}$	**0.988** DF-500-Dice-5	0.986 CHI-500-Dice-3 CHI-500-Dice-5 CHI-500-Dice-7	0.986 CHI-500-Dice-5 CHI-500-Dice-7	0.989 DF-500-Dice-3 DF-500-Dice-5 DF-500-Dice-7	0.991 DF-500-Dice-5
●●○$_{LC✓ST✓SW*}$	0.986 CHI-500-Dice-5	0.986 CHI-250-Dice-5	0.983 CHI-250-Dice-3 CHI-500-Dice-3 CHI-500-Dice-7 DF-500-Dice-7	0.989 CHI-250-Dice-5 CHI-500-Dice-5 CHI-500-Dice-7	0.989 CHI-500-Dice-5
●●●$_{LC✓ST✓SW✓}$	0.985 CHI-250-Dice-3 CHI-500-Dice-7 DF-500-Dice-5	0.982 CHI-250-Dice-3 CHI-250-Dice-5	0.984 CHI-500-Dice-7	0.991 DF-500-Dice-7	**0.992** CHI-250-Dice-3

Note: The bold and underlined value indicates the maximum preprocessing value in the columns.

Table 6. Maximum Micro-F1 scores of E-Ds and the methods utilized

	E-Ds1	**E-Ds2**	**E-Ds3**	**E-Ds4**	**E-Ds5**
○○○$_{LC*ST*SW*}$	0.953 OR-500-Dice-3	0.945 OR-250-Dice-5	0.942 OR-500-Inner-3	0.938 CHI-250-Inner-3 OR-500-Inner-3	0.923 OR-250-Dice-3
○○●$_{LC*ST*SW✓}$	**0.959** CHI-500-Dice-5	0.957 CHI-500-Dice-3	0.954 OR-500-Dice-7	0.953 CHI-500-Dice-3 CHI-500-Dice-7	0.929 OR-500-Dice-3
○●○$_{LC*ST✓SW*}$	0.950 OR-500-Inner-3	0.949 OR-500-Inner-3	0.946 OR-500-Inner-3	0.947 OR-500-Inner-3	**0.939** OR-250-Dice-3
○●●$_{LC*ST✓SW✓}$	0.952 OR-500-Dice-5	0.955 CHI-250-Dice-3 OR-500-Dice-3 OR-500-Dice-7	0.947 CHI-500-Dice-7 OR-500-Dice-7	0.949 OR-500-Dice-3	0.935 CHI-250-Dice-3
●○○$_{LC✓ST*SW*}$	0.951 OR-500-Inner-3	0.951 OR-250-Dice-7	0.956 OR-500-Dice-3	0.948 OR-500-Inner-3	0.938 OR-500-Dice-3
●○●$_{LC✓ST*SW✓}$	0.956 OR-500-Dice-5	**0.964** CHI-250-Dice-3 CHI-250-Dice-5	0.955 OR-500-Dice-3	0.955 CHI-500-Dice-3	**0.939** CHI-500-Dice-3
●●○$_{LC✓ST✓SW*}$	0.943 OR-500-Inner-3	0.951 OR-500-Inner-3	0.947 CHI-250-Dice-7	0.947 CHI-100-Dice-5	0.934 OR-250-Dice-5
●●●$_{LC✓ST✓SW✓}$	0.952 OR-500-Dice-3	0.958 CHI-250-Dice-3	**0.957** OR-500-Dice-3	**0.956** CHI-500-Dice-3	**0.939** CHI-250-Dice-3

Note: The bold and underlined value indicates the maximum preprocessing value in the columns.

Table 7. Average Micro-F1 scores of preprocessing methods in T-Ds

	T-Ds1	T-Ds2	T-Ds3	T-Ds4	T-Ds5	
○○○$_{LC*ST*SW*}$	0.851	0.860	0.879	0.889	0.903	0.876
○○●$_{LC*ST*SW✓}$	0.842	0.852	0.872	0.875	0.916	0.871
○●○$_{LC*ST✓SW*}$	0.915	0.914	0.924	0.933	0.933	0.924
○●●$_{LC*ST✓SW✓}$	**0.921**	**0.918**	**0.929**	**0.937**	**0.943**	**0.930**
●○○$_{LC✓ST*SW*}$	0.864	0.870	0.883	0.901	0.909	0.885
●○●$_{LC✓ST*SW✓}$	0.855	0.859	0.880	0.889	0.922	0.881
●●○$_{LC✓ST✓SW*}$	0.910	0.909	0.919	0.929	0.930	0.919
●●●$_{LC✓ST✓SW✓}$	0.916	0.915	0.925	0.932	0.941	0.926
	0.884	0.887	0.901	0.911	0.925	0.902

Note: The bold and underlined value indicates the maximum preprocessing value in the columns.

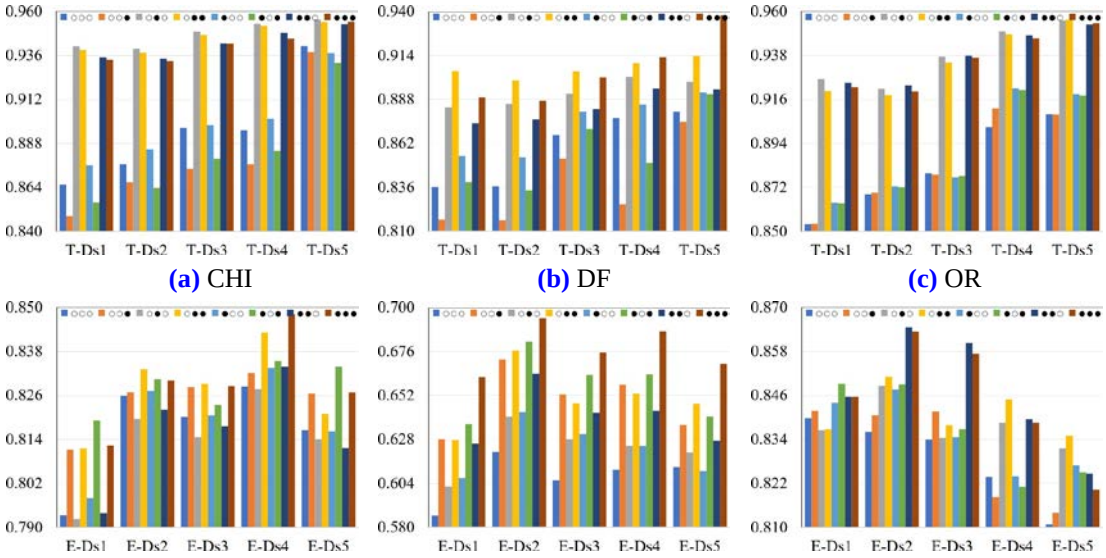
Table 8. Average Micro-F1 scores of preprocessing methods in E-Ds

	E-Ds1	E-Ds2	E-Ds3	E-Ds4	E-Ds5	
○○○$_{LC*ST*SW*}$	0.730	0.752	0.742	0.742	0.742	0.742
○○●$_{LC*ST*SW✓}$	0.770	0.789	0.784	0.781	0.763	0.777
○●○$_{LC*ST✓SW*}$	0.737	0.762	0.751	0.754	0.753	0.752
○●●$_{LC*ST✓SW✓}$	0.765	0.793	0.780	0.790	0.770	0.780
●○○$_{LC✓ST*SW*}$	0.741	0.764	0.752	0.749	0.747	0.751
●○●$_{LC✓ST*SW✓}$	**0.778**	0.795	0.784	0.784	**0.771**	0.782
●●○$_{LC✓ST✓SW*}$	0.749	0.777	0.766	0.763	0.752	0.761
●●●$_{LC✓ST✓SW✓}$	0.774	**0.800**	**0.790**	**0.792**	0.769	**0.785**
	0.755	0.779	0.769	0.770	0.758	0.766

Note: The bold and underlined value indicates the maximum preprocessing value in the columns.

The average scores of the components related to the classification used in the study are given in **Fig. 2** - **Fig. 4** to display whether the preprocessings affect the classification performance of feature selection, feature size and classifier, and the direction of the effects if they affect. The following results were obtained: the lowercase conversion (●--$_{LC✓}$) in T-Ds{1-5} datasets does not have a significant effect on classification success; the classification success is increased generally when CHI, DF, OR, 100, 250, 500, Dice and Euclid are applied with stemming (-●-$_{ST✓}$); removal of stop words (--●$_{SW✓}$) in CHI, OR, Euclid has generally a negative effect on success and has generally a positive effect on success in 100, Dice, Inner, MNB. The following results were obtained: the lowercase conversion (●--$_{LC✓}$) in E-Ds{1-5} datasets does not have a significant effect on classification success as is the case in T-Ds datasets; the classification success is increased generally when DF, OR, 250, Euclid, SVM are applied with stemming (-●-$_{ST✓}$) but decreased in Inner; the classification success is increased generally when the stop words are removed (--●$_{SW✓}$) in CHI, DF, 100, 250, 500, Dice, Inner, SVM but generally decreased in Euclid.

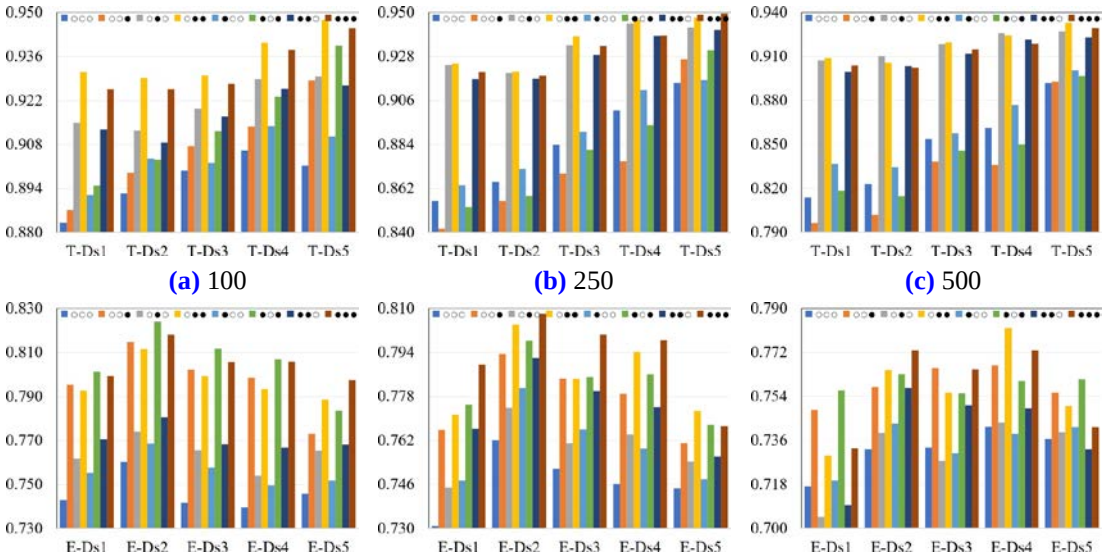
In summary, different results were obtained in the average Micro-F1 scores of preprocessing and other classification components in T-Ds and E-Ds datasets. Furthermore, the results displayed minor or major differences according to the language, the situation whether the dataset is balanced or not and even to imbalance level, used methods.



(a) CHI **(b)** DF **(c)** OR

(d) CHI **(e)** DF **(f)** OR

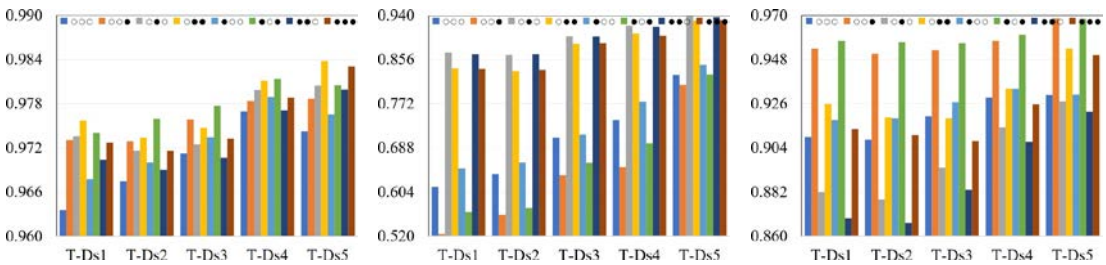
Fig. 2. Average Micro-F1 scores of preprocessings with feature selection methods in T-Ds (a - c) and E-Ds (d - f)



(a) 100 **(b)** 250 **(c)** 500

(d) 100 **(e)** 250 **(f)** 500

Fig. 3. Average Micro-F1 scores of preprocessings with feature sizes in T-Ds (a - c) and E-Ds (d - f)



(a) Dice **(b)** Euclid **(c)** Inner

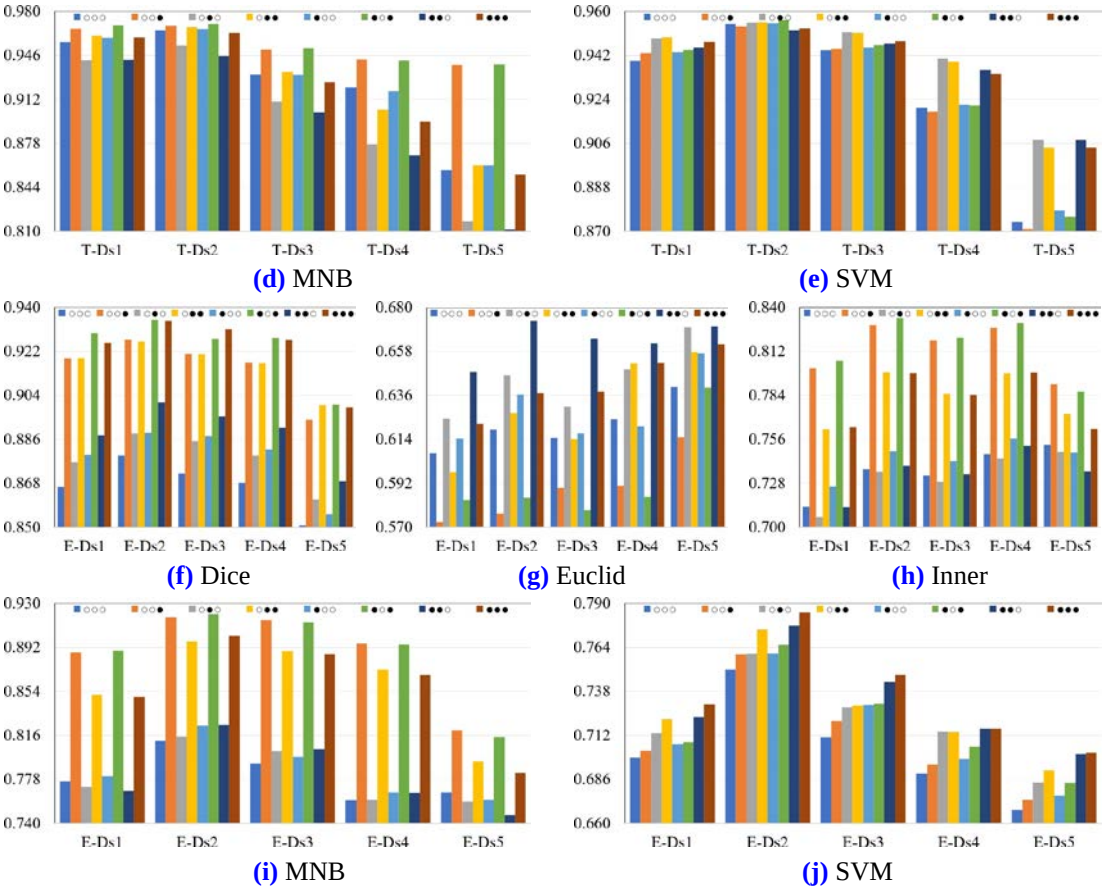


Fig. 4. Average Micro-F1 scores of preprocessings with classifiers in T-Ds (a - e) and E-Ds (f - j)

3.1 Comparative Analysis

In **Table 9**, information about dataset count, whether the dataset is balanced or imbalanced, which of the preprocessing methods were used or not used, feature selection, feature size, classification algorithms and maximum scores obtained from datasets, used in our study and studies in the literature are presented.

There are studies in which the effects of the factors affecting the success of text classification are analyzed. In the review, in other studies, it was observed that Zemberek stemmer [14, 15, 18, 28] and Porter Stemmer [14, 20, 28] were preferred for stemming; Chi-Square [14, 28] and Odds Ratio [14] for feature selection and kNN [5, 17, 24, 25], MNB [18, 20] and SVM [5, 14, 17, 18, 20, 25, 27, 28] for feature selection as in this study.

Preprocessing, which is the focus of this study, has been defined in a variety of ways in the literature, such as the fact that it is one of the most crucial and fundamental aspects of text classification [14, 16, 19, 28], removes unnecessary data before analysis [5] and is used to enhance the quality of data in the classic machine learning algorithm for text classification [18]. It can be said that lowercase conversion, stemming, and stop word removal [5, 14-18, 28] are the most popular preprocessing techniques, even though the same preprocessing is not used in every classification study. As also obtained from our study; it was reported in studies that there is no certain preprocessing method that is successful in all languages [28], preprocessing is necessary for text classification, it increases the classification success, the effect of each preprocessing is not the same, and the preprocessing steps differ from language to language

[14, 15] and classifiers are affected from preprocessing [14]. However, there is also study expressing that preprocessing methods do not have a significant effect on classification [5, 14]. According to the findings obtained within the scope of our study, it should be noted that this is true in terms of the highest Micro-F1 scores, but not true in terms of the average Micro-F1 scores.

It has been emphasized that when the results are examined in terms of preprocessing techniques, it is impossible to enunciate whether the stop words removal and lemmatization (stemming) processes have a positive or negative impact on the results because this may vary depending on the language, classifier, and dataset [17]. It has been explained that the appropriate combination of preprocessings performed depending on the language will increase the success, otherwise, it may decrease it, however, the preprocessing steps are as important as feature extraction, feature selection and classifier [28]. While it was expressed that stemming in one dataset increases the classification, although immediately not apparent, in another dataset, more successful classifications are made with lowercase conversion and stop words removal, and combination preprocessing methods are more successful than singles and preprocessing effects may differ according to the dataset [5], whereas in another study it was stated stemming reduces success [16].

As in our study, the most effectual preprocessing method for reducing the number of discrete words for Turkish and English is stemming [14]. Again, in our study, the highest classification was acquired for Turkish as 0.993 and for English as 0.964. In studies in the literature, it is observed that the success is 0.781 [14], 0.920 [18] and 0.935 [15] in Turkish datasets, and 0.799 [16], 0.961 [5] and 0.980 [14] in datasets in other languages. It was observed that the highest Turkish news classification was obtained with alphabetic tokenization and when lowercase conversion and stemming are applied, and stop words removal is not applied with SVM as 0.806 [28]; in another study in two different datasets, among classic machine learning algorithms with SVM as 0.952 and 0.918, and in Pre-trained language models with BERT model as 0.963 and 0.960 was achieved [17]; and in another study it was acquired as 0.781 with SVM when stop words removal was applied but stemming was not [14]. It was determined that the highest English news classification with lowercase conversion and stemming are applied, and alphabetic tokenization and stop words removal is not applied with SVM as 0.872 [28]; in another study obtained as 0.980 with Decision Tree when stemming was applied and stop words removal was not [14]. When compared these values to other studies, our classifications are more successful for Turkish classification and closely align with English classification. As in our study, it was observed that preprocessing did not have a positive effect on Turkish, but it had a significant contribution on English [14]. On the other hand, research has shown that updated/developed methods outperform classic machine learning algorithms [5, 17, 18]. To determine whether or which preprocessing method(s) is required, consideration should also be given to the language, document type, feature selection, feature size, and classifier used. The dataset does, however, have an impact on the classification outcomes, and this should not be disregarded. Although the focus of this study is to investigate the effects of preprocessing on balanced and imbalanced datasets in text classification, it could be useful to examine the effects of preprocessing and other factors affecting classification, especially on datasets obtained from different domains in future studies.

Table 9. Comparison this study with the literature

Study	Dataset Count	Balanced Or Imbalanced	LC	ST	SW	Feature Selection	Feature Size	Classifier	Maximum Score
[15]	1 Turkish	Balanced	✓	✓	✓	Correlation-based, attribute ranking-based	No information	Cosine similarity	Turkish: 0.935
[14]	1 Turkish, 1 English	Balanced	✓	Both	Both	Gini Index, Normalized Difference Measure, Extensive Feature Selector	50, 100, 300, 500, 1000	J48 Decision Tree, SVM	Turkish: 0.781 English: 0.980
[17]	2 Turkish	Balanced	Both	Both	Both	No information	100, 500, and 1000 to 500000 increment 1000	Supervised machine learning algorithms, Pre-trained language models	Turkish1: 0.963 Turkish2: 0.960
[18]	2 Turkish	Augmented/balanced, Imbalanced	✓	✓	✓	No information	No information	Classical machine learning models, BERT models	Turkish1: 0.920 Turkish2: 0.920
[5]	3 English	Binary/imbalanced	Both	Both	Both	TF-IDF, PCA, fuzzy matching, Euclidean distance similarity	No information	Naive Bayes, kNN, Deep learning	English1: 0.957 English2: 0.961 English3: 0.956
[28]	2 Turkish, 2 English	Binary/balanced, Multiclass/imbalanced	Both	Both	Both	Chi-square	10, 20, 50, 100, 200, 500, 100, 2000	SVM	Turkish1: 0.971 Turkish2: 0.806 English1: 0.989 English2: 0.872
This Study	*5 Turkish, 5 English*	*1 balanced and 4 imbalanced datasets for Turkish and English*	*Both*	*Both*	*Both*	*Chi-square, Document Frequency, Odds Ratio*	*100, 250, 500*	*Dice Coefficient, Euclid, Inner Product, MNB, SVM*	*T-Ds1: 0.988* *T-Ds2: 0.988* *T-Ds3: 0.989* *T-Ds4: 0.993* *T-Ds5: 0.992* *E-Ds1: 0.959* *E-Ds2: 0.964* *E-Ds3: 0.957* *E-Ds4: 0.956* *E-Ds5: 0.939*

4. Conclusion

The effects of preprocessing methods, which are tested in balanced and imbalanced datasets in Turkish and English languages, on text classification were examined in this study. These effects were reviewed with the sub-dimensions of word number, classification success, language, processing time and whether the dataset is balanced or not. Moreover, the effects of preprocessing on feature selection, feature size and classification algorithms, which are other factors of affecting classification, were also examined comparatively.

When the preprocessings are considered as separately as ○○●$_{LC\times ST\times SW\checkmark}$, ○●○$_{LC\times ST\checkmark SW\times}$ and ●○○$_{LC\checkmark ST\times SW\times}$, albeit at different rates the changes in the number of discrete and total words were obtained with the same processes in Turkish and English datasets. The highest decrease was obtained by applying stemming in the number of discrete word and removing stop words in the number of total words.

According to maximum Micro-F1 values, the following outcomes were observed: different results are obtained between languages, Turkish datasets are not much affected by the applied preprocessing methods and English datasets were relatively more affected. Changes in all these processes are below 1% in Turkish datasets (between the intervals of ‰3 and ‰8); around 2% in English datasets (between the intervals of ‰16 and ‰20). Whether the dataset was balanced or imbalanced affected the Turkish classification by 1% and the English classification by 4%.

According to the average Micro-F1 scores, the highest success in Turkish datasets was obtained by applying stemming and removing stop words (○●●$_{LC\times ST\checkmark SW\checkmark}$; 0.930), and the classification success increased as the imbalance rate of the dataset increased. On the other hand, the highest average success in English datasets was obtained by applying lowercase and stemming, removing stop words (●●●$_{LC\checkmark ST\checkmark SW\checkmark}$; 0.785) and as the imbalance rate of the dataset increased, the classification success mostly decreased. It was seen that the stemming application is the most determining preprocessing method which cause increase in success on text classification carried out with Turkish documents and stop words removal on English text classification. Furthermore, it was determined that success generally increases with preprocessing combinations in English datasets, and also it was seen that binary preprocessings are more successful than singles (○●●$_{LC\times ST\checkmark SW\checkmark}$ than ○○●$_{LC\times ST\times SW\checkmark}$ and ○●○$_{LC\times ST\checkmark SW\times}$; ●○●$_{LC\checkmark ST\times SW\checkmark}$ than ○○●$_{LC\times ST\times SW\checkmark}$ and ●○○$_{LC\checkmark ST\times SW\times}$; ●●○$_{LC\checkmark ST\checkmark SW\times}$ than ○●○$_{LC\times ST\checkmark SW\times}$ and ●○○$_{LC\checkmark ST\times SW\times}$) and triplets are more successful than doubles (●●●$_{LC\checkmark ST\checkmark SW\checkmark}$; than ○●●$_{LC\times ST\checkmark SW\checkmark}$, ●○●$_{LC\checkmark ST\times SW\checkmark}$ and ●●○$_{LC\checkmark ST\checkmark SW\times}$).

Different results were observed in Turkish and English documents in the classifications of classes with few numbers of documents. In Turkish datasets, it was observed that about half of the documents in the class are classified correctly, and the misclassifications are assigned to class that is close to the related class in terms of subject. In English datasets, it was determined that approximately 30% of the documents in class are classified correctly, and the misclassifications are done in the direction of the class with the highest document.

The preprocessing methods, in which the processing time and the highest decrease in the total number of words occur, are the same and differences between languages were seen. Processes were completed in a shorter time with stemming in Turkish datasets and stop words removal in English datasets.

It was seen that the performances of feature selection, feature size and classifier generally increase by applying stemming in Turkish datasets and stop words removal in English datasets. Moreover, it was concluded that not only preprocessing but also feature selection, feature size and classifier affect the classification success. Furthermore, it was observed that feature selection, feature size and classifier are more determining dominantly more effective than preprocessing in classification.

As a result, by considering the highest Micro-F1 values, it can be concluded that there is no need to apply preprocessing, and that classification can be made with high success without applying preprocessing. However, when preprocessing is applied, it was determined that there is an improvement in processing times; for example, it was found out that processing times are reduced by almost 90% with stemming in Turkish datasets and with stop words removal in English datasets. When the average Micro-F1 scores are examined, the importance, necessity and effects of preprocessing for classification, and the fact that preprocessing methods produce different results from language to language can be seen more clearly. Considering the advantages it provides in terms of processing time, as well as its effects on text classification, preprocessing is necessary, since the dynamics of languages are different, it would be more accurate to consider and evaluate each language separately, different preprocessing process(es) would be more appropriate for different languages, a preprocessing method may not produce the same results in all languages, preprocessing that is successful in one language may not have the same effect in another language, and even a preprocessing that is successful in one language can have a negative effect in another language.

References

[1] A. Gupta and R. Bhatia, "Ensemble approach for web page classification," *Multimedia Tools and Application*, vol. 80, pp. 25219-25240, 2021. Article (CrossRef Link)

[2] E. Buber and B. Diri, "Web page classification using RNN," *Procedia Computer Science*, vol. 154, pp. 62-72, 2019. Article (CrossRef Link)

[3] S. Seo, C. Kim, H. Kim, K. Mo, and P. Kang, "Comparative study of deep learning-based sentiment classification," *IEEE Access*, vol. 8, pp. 6861-6875, 2020. Article (CrossRef Link)

[4] M. Y. Pak and S. Gunal, "Sentiment classification based on domain prediction," *Elektronika Ir Elektrotechnika*, vol. 22, no. 2, pp. 96-99, 2016. Article (CrossRef Link)

[5] V. Nurcahyawati and Z. Mustaffa, "Improving sentiment reviews classification performance using support vector machine-fuzzy matching algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1817-1824, 2023. Article (CrossRef Link)

[6] S. Khedkar and S. Shinde, "Deep learning and ensemble approach for praise or complaint classification," *Procedia Computer Science*, vol. 167, pp. 449-458, 2020. Article (CrossRef Link)

[7] Y. Yang, D. -L. Xu, J.-B. Yang, and Y. -W. Chen, "An evidential reasoning-based decision support system for handling customer complaints in mobile telecommunications," *Knowledge-Based Systems*, vol. 162, pp. 202-210, 2018. Article (CrossRef Link)

[8] B. K. Dedeturk and B. Akay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm," *Applied Soft Computin*, vol. 91, 106229, 2020. Article (CrossRef Link)

[9] N. Saidani, K. Adi, and M. S. Allili, "A semantic-based classification approach for an enhanced spam detection," *Computers & Security*, vol. 94, 101716, 2020. Article (CrossRef Link)

[10] M. Duşçu and D. Günneç, "Polarity classification of Twitter messages using audio processing," *Information Processing & Management*, vol. 57, no. 6, 102346, 2020. Article (CrossRef Link)

[11] S. M. Khan, M. Chowdhury, L. B. Ngo, and A. Apon, "Multi-class twitter data categorization and geocoding with a novel computing framework," *Cities*, vol. 96, 102410, 2020. Article (CrossRef Link)

[12] A. Devaraj, D. Murthy, and A. Dontula, "Machine-learning methods for identifying social media-based requests for urgent help during hurricanes," *International Journal of Disaster Risk Reduction*, vol. 51, 101757, 2020. Article (CrossRef Link)

[13] C. I. Hausladen, M. H. Schubert, and E. Ash, "Text classification of ideological direction in judicial opinions," *International Review of Law and Economics*, vol. 62, 105903, 2020. Article (CrossRef Link)

[14] B. Parlak, "The effects of preprocessing on Turkish and English news data," *Sakarya University Journal of Computer and Information Sciences*, vol. 6, no. 1, pp. 59-66, 2023. Article (CrossRef Link)

[15] A. Yürekli, "On the effectiveness of paragraph vector models in document similarity estimation for Turkish news categorization," *Eskişehir Technical University Journal of Science and Technology A - Applied Sciences and Engineering*, vol. 24, no. 1, pp. 23-34, 2023. Article (CrossRef Link)

[16] G. Jung, J. Shin, and S. Lee, "Impact of preprocessing and word embedding on extreme multi-label patent classification tasks," *Applied Intelligence*, vol. 53, 4047-4062, 2023. Article (CrossRef Link)

[17] Ö. Köksal and E. H. Yılmaz, "Improving automated Turkish text classification with learning-based algorithms," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 11, p. e6874, 2022. Article (CrossRef Link)

[18] P. Uskaner Hepsağ, S. A. Özel, K. Dalcı, and A. Yazıcı, "Using BERT models for breast cancer diagnosis from Turkish radiology reports," *Language Resources and Evaluation*, 2023. Article (CrossRef Link)

[19] 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. Article (CrossRef Link)

[20] L. Chen, L. Jiang, and C. Li, "Using modified term frequency to improve term weighting for text classification," *Engineering Applications of Artificial Intelligence*, vol. 101, p. 104215, 2021. Article (CrossRef Link)
[21] D. Rohidin, N. A. Samsudin, and M. M. Deris, "Association rules of fuzzy soft set based classification for text classification problem," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 801-812, 2022. Article (CrossRef Link)
[22] Ö. Köksal, "Tuning the Turkish text classification process using supervised machine learning-based algorithms," in *Proc. of 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, Novi Sad, Serbia, pp. 1-7, 2020. Article (CrossRef Link)
[23] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 2, pp. 225-231, 2020. Article (CrossRef Link)
[24] T. Dogan and A. K. Uysal, "A novel term weighting scheme for text classification: TF-MONO," *Journal of Informetrics*, vol. 14, no. 4, 101076, 2020. Article (CrossRef Link)
[25] T. Dogan and A. K. Uysal, "Improved inverse gravity moment term weighting for text classification," *Expert Systems with Applications*, vol. 130, pp. 45-59, 2019. Article (CrossRef Link)
[26] M. U. Salur and İ. Aydın, "The impact of preprocessing on classification performance in convolutional neural networks for Turkish text," in *Proc. of 2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, pp. 1-4, 2018. Article (CrossRef Link)
[27] A. Rehman, K. Javed, and H. A. Babri, "Feature selection based on a normalized difference measure for text classification," *Information Processing and Management*, vol. 53, no. 2, pp. 473-489, 2017. Article (CrossRef Link)
[28] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Information Processing and Management*, vol. 50, no. 1, pp. 104-112, 2014. Article (CrossRef Link)
[29] T. Parlar and E. Sarac, "IWD based feature selection algorithm for sentiment analysis," *Elektronika Ir Elektrotechnika*, vol. 25, no. 1, pp. 54-58, 2019. Article (CrossRef Link)
[30] Zemberek, "An open source NLP library for Turkic languages," 2021. [Online]. Available: http://code.google.com/p/zemberek
[31] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130-137, 1980.

Mehmet F. Karaca received his MSc and PhD in Computer Engineering from Karabuk University, Turkey in 2012 and 2018, respectively. He is an Assistant Professor at the Department of Management Information Systems in Erbaa Faculty of Social Sciences and Humanities, Tokat Gaziosmanpaşa University, Erbaa, Tokat, Turkey. His current research interests include Natural Language Processing, Data Mining, Human Computer Interaction and Turkish Sign Language.