Computer Science • 20(1) 2019 https://doi.org/10.7494/csci.2019.20.1.3097

Tuba Parlar
Selma Ayşe Özel
Fei Song

# ANALYSIS OF DATA PRE-PROCESSING METHODS FOR SENTIMENT ANALYSIS OF REVIEWS



**Abstract** *The goals of this study are to analyze the effects of data pre-processing methods for sentiment analysis and determine which of these pre-processing methods (and their combinations) are effective for English as well as for an agglutinative language like Turkish. We also try to answer the research question of whether there are any differences between agglutinative and non-agglutinative languages in terms of pre-processing methods for sentiment analysis. We find that the performance results for the English reviews are generally higher than those for the Turkish reviews due to the differences between the two languages in terms of vocabularies, writing styles, and agglutinative property of the Turkish language.*

**Keywords** data pre-processing, feature selection, sentiment analysis, text classification

**Citation** Computer Science 20(1) 2019: 123–141

## 1. Introduction

Also known as opinion mining, sentiment analysis is a natural language processing task that tries to extract sentiment-expressing features and determine the polarity of a review document as "positive," "negative", or occasionally "neutral". A sentiment analysis for a review document captures the author's opinion, judgment, or emotion of the entities covered in the text. Therefore, it can have many useful applications like opinionated web searches and the automatic analysis of reviews [22]. In particular, sentiment analysis allows us to know what other people think about such entities as products, services, and companies so we can make informed decisions.

Sentiment analysis is essentially a text-classification process, since the main steps (like data pre-processing, feature selection, and classification) are also applied to sentiment analysis. However, sentiment classification is different from the traditional topic-based classification in that it requires different techniques to select sentiment-expression features in order to sort the review documents into different polarities.

Studies about sentiment analysis have been increased recently due to its wide range of applications in business. However, most of these studies focus on the use of different approaches for feature selection and classification [1, 2, 12, 18, 20, 22–24, 27, 28]. There is still a lack of comprehensive studies on data pre-processing methods, particularly for different languages [1, 5, 14, 19, 24, 28]. In this study, our aim is to investigate the effects of data pre-processing methods (including stemming, stop word removal, and punctuation removal) on the accuracy of sentiment analysis for review documents. In the topical classification of text, stemming as well as the removal of stop words and punctuation marks are usually applied to reduce the feature size and improve the classification accuracy. However, punctuation marks and stop words may be important in sentiment analysis, as they can be used to express sentiments. We also want to study the effects of stemming for review documents from English and an agglutinative language like Turkish. To our knowledge, there is currently no study that performs a detailed analysis of the pre-processing methods for Turkish and compares the results with the English results. Therefore, our study should be useful for researchers who work on the text processing of agglutinative languages.

The Turkish language belongs to the Altaic branch of the Ural-Altaic family of languages and is mainly used in the Republic of Turkey. The Turkish alphabet is based on Latin characters and has 29 letters, consisting of 8 vowels (a, e, ı, i, o, ö, u, ü) and 21 consonants (b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z). Turkish is an agglutinative language similar to Finnish and Hungarian, where a single word can be translated into a relatively longer sentence in English (as shown in Table 1 [9]). In such a morphologically rich language, the structure of a word carries a lot of information such as the part of speech, modality, tense, person and number agreement, polarity, case, and voice. In addition, a new word can be formed by adding a suffix to a stem. For example, göz (meaning "eye") + -1IK = gözlük ("spectacles") + -CI = gözlükçü ("optician"), or göz ("eye") + -CI = gözcü ("guard") + -1IK = gözcülük ("being a guard"). Some suffixes can also appear recursively in

the same word: e.g., göz-lük-cü-lük ("the profession/business of an optician"). Thus, when stemming is applied, the meaning of a word can be changed, which can decrease the performance of a sentiment analysis of Turkish reviews.

**Table 1**
Two example single-word sentences in Turkish with English translations

| Turkish | Morphological Analysis | English Meaning |
|---|---|---|
| Duyamazsın | hear-Caus-Abil-Neg-Pres-P2sg | (You) can/may not hear |
| Karşılaştırmalısın | compare-Caus-Oblig-P2sg | (You) must make (something) compare |

In this study, we use three different supervised machine-learning techniques: Naïve Bayes Multinomial (NBM), Support Vector Machines (SVM), and C4.5 decision tree classifiers to examine the effects of pre-processing methods for sentiment analysis. NBM is based on probabilities, while SVM is based on data selection – both of which are widely used for sentiment analysis [2, 12, 23, 27]. C4.5 is another classifier that is especially used for web document classification [18, 23, 27]. Movie and product review data sets (namely, books, DVDs, electronics, and kitchen appliances) in both the English and Turkish languages are used in our experiments. We also apply our proposed feature selection method (QER), which was especially developed for sentiment analysis to remove irrelevant and noisy features and to improve classification accuracy [24].

The remainder of this paper is organized as follows: Section 2 summarizes related works and lists the contributions of this study. Section 3 describes the data sets, classifiers, and feature-selection methods used to evaluate the data pre-processing methods. Section 4 discusses the experimental results, and Section 5 ultimately concludes the paper.

## 2. Related Work

There are a number of studies about sentiment analysis that have used different approaches for data pre-processing, feature vector construction, feature selection, and classification [1, 2, 18, 22–24, 27, 28]. Data pre-processing may contain such tasks as punctuation removal, case normalization, stop word removal, and stemming. Feature vector construction commonly uses bags-of-words and represents features such as weighted vectors for documents. In the classification phase, supervised or unsupervised methods are used. In the supervised methods, a classifier is trained by documents with known class labels. After that, the classifier is used to predict the classes of test documents that have yet to be seen. The unsupervised methods use a lexicon to compute the polarities of texts.

Most researchers apply supervised learning techniques because they can be automatically trained and improved through training data sets. Pang, Lee, and

Vaithyanathan [23] implement supervised text classification methods (namely, Naïve Bayes (NB), SVM, and Maximum Entropy Modeling (MEM) classifiers) and select features based on n-grams and different term-weighting methods: term frequency and term presence weighting. They find that uni-grams with the term presence weighting method perform better than the other combinations. According to the experiments performed on movie reviews, they find that SVM performs the best, with an accuracy of 82.9% using three-fold cross validation.

Some researchers investigate the effects of pre-processing methods for sentiment analysis. For example, Duwairi and El-Orfali [14] examine the effects of pre-processing methods for Arabic reviews. They evaluate the performance of their pre-processing methods using three text classifiers: NB, SVM, and KNN (K-Nearest Neighbors).

Sentiment analysis has been studied extensively, but most of the works have been specific to the English language. Recently, sentiment analysis has also become active for Turkish researchers. Table 2 gives a summary of the related works for Turkish reviews. As can be seen, only a few researchers investigate the effects of data pre-processing for the sentiment analysis of Turkish reviews. As Turkish is an agglutinative language, the related works listed in Table 2 consider the effects of stemming only. When extracting features from the data set, it is a common practice to eliminate all punctuation marks [15]. However, as pointed out by Devitt and Ahmad [13], certain textual features such as lexical, syntactic, and punctuation marks also contribute to the meaning of the text.

**Table 2**
Summary of related work on the sentiment analysis of Turkish reviews

| References | Data Scope | Pre-processing | Text Classification |
|---|---|---|---|
| Erogul [15] | movie | Stemming (on/off) | SVM |
| Kaya et al. [16] | political news | Stemming (on/off) | NB, ME, SVM |
| Cetin et al. [10] | tweets | Stemming (on/off) | NB, CNB, SVM |
| Demirtas et al. [11] | movie, product | Stemming (off) | NB, SVM |
| Sevindi [25] | movie | Stemming (on/off)<br>Stopwords(on/off) | NB, KNN, C4.5 |
| Akba et al. [3] | movie | Stemming (on) | NB, SVM |

To the best of our knowledge, only Kaya et al. [16] study the impacts of keeping punctuation marks as a textual feature in Turkish reviews. Some researchers remove all of the stop words to reduce the feature sizes [15]. However, Kaya et al. [16] choose to keep stop words because they believe that many stop words can carry sentiments. Also, stemming is used to reduce the feature sizes in most of the research [3, 10, 11, 15, 16, 21, 25]. Erogul [15] examines the effects of stemming while using n-grams as features and SVM as the classifier. He obtains the maximum $F$-measure (85%) with a spellchecking method without stemming. Kaya et al. [16] also examine the use of the roots of words as features with NB, SVM, and MEM classifiers but find that there is no significant differences in the results. Sevindi [25] investigates

the effects of stemming using NB, SVM, C4.5, and KNN, but the performance is decreased.

As can be seen in Table 2, previous researchers have investigated the effects of stemming and stop word removal separately, and they generally removed the punctuation marks. The contributions of our study can be summarized as follows: we analyze the combined effects of stemming, stop word removal, and punctuation marks for both Turkish and English reviews and try to determine which pre-processing combinations should be used for the sentiment analysis of Turkish and English reviews. To reach our goal, we consider all possible combinations of the options for data pre-processing and test them against multiple data sets (such as movie reviews and product reviews) in both Turkish and English using three machine-learning algorithms so we can examine their interactions with different classifiers and languages. We also apply our proposed feature-selection method and list the most valuable punctuation marks and stop words for the Turkish and English data sets that can be useful for sentiment analysis.

## 3. Materials and Methods

In this section, we explain the data sets, pre-processing methods, classifiers, and performance evaluation measures used for our experiments in detail.

### 3.1. Data sets

We use movie and product review datasets in Turkish and English. The Turkish movie review dataset [25] is collected from beyazperde.com and is comprised of 1057 positive and 978 negative reviews. The Turkish product review datasets are collected from hepsiburada.com by Demirtas and Pechenizky [11]. They consist of four categories about books, DVDs, electronics, and kitchen appliances, and each category has 700 positive and 700 negative reviews. Table 3 summarizes the general statistics of these five Turkish review datasets.

The English movie review dataset is introduced by Pang and Lee [21] and is comprised of 1000 positive and 1000 negative reviews collected from rottentomatoes.com. The English product review datasets are collected from amazon.com by Blitzer et al. [7] in four categories: books, DVDs, electronics, and kitchen appliances, each consisting of 1000 positive and 1000 negative reviews. Table 4 summarizes the general statistics of these five English review datasets.

In Tables 3 and 4, we can observe that the numbers of samples in the English and Turkish datasets are almost the same; however, the numbers of words in the Turkish datasets are fewer than in the English datasets, as Turkish is an agglutinative language where one word can express one sentence in English. Therefore, the average characters per word in the Turkish data sets are greater than those in the English data sets.

**Table 3**
General statistics of Turkish review data sets

| | Movie Reviews | DVDs Reviews | Electronics Reviews | Book Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| Number of reviews | 2035 | 1400 | 1400 | 1400 | 1400 |
| Total words | 80,777 | 45,469 | 52,716 | 47,647 | 46,273 |
| Avg. characters/word | 5.92 | 5.99 | 6.0 | 6.06 | 6.0 |
| Avg. words/sentence | 16.92 | 14.18 | 13.61 | 14.07 | 13.0 |
| Avg. words/document | 39.69 | 32.48 | 37.65 | 34.03 | 33.05 |

**Table 4**
General statistics of English review data sets

| | Movie Reviews | DVDs Reviews | Electronics Reviews | Book Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| Number of reviews | 2000 | 1400 | 1400 | 1400 | 1400 |
| Total words | 1,329,753 | 238,192 | 155,422 | 246,387 | 131,814 |
| Avg. characters/word | 4.48 | 4.32 | 4.19 | 4.48 | 4.13 |
| Avg. words/sentence | 18.59 | 18.12 | 15.71 | 18.68 | 15.03 |
| Avg. words/document | 664.88 | 170.14 | 111.02 | 175.99 | 94.15 |

## 3.2. Data Pre-processing Methods

Data pre-processing in sentiment analysis is the process of preparing the text for classification. In this study, the pre-processing steps include tokenization, punctuation removal, stop word removal, stemming, and document vector construction. Tokenization is a crucial procedure of splitting a text into meaningful units called tokens. For each token obtained, we apply case normalization; then, we consider whether to keep the punctuation marks or not, remove the stop words or not, and perform the stemming or not, giving us a total of eight combinations for the data pre-processing. For punctuation marks, we identify a total of 13 patterns that may be useful for sentiment analysis, which are summarized in Table 5 (along with explanations and matched examples). Other punctuation patterns are eliminated to reduce our feature size.

**Table 5**
Punctuation patterns used in our experiments

| # | Explanations | Examples |
|---|---|---|
| 1. | All tokens starting with colon and one or more closed parentheses | :) :)) :))) |
| 2. | All tokens starting with colon and one or more open parentheses | :( :(( :((( |
| 3. | All tokens starting with equal sign and one or more closed parentheses | =) =)) =))) |
| 4. | All tokens starting with equal sign and one or more open parentheses | =( =(( =((( |
| 5. | Sequences of at least two periods | .. ... .... |
| 6. | Sequences of at least two exclamation marks | !! !!! !!!! |
| 7. | Sequences of at least two question marks | ?? ??? ???? |
| 8. | Single period | . |
| 9. | Single comma | , |
| 10. | Single exclamation mark | ! |
| 11. | Single question mark | ? |
| 12. | Single colon | : |
| 13. | Single semicolon | ; |

For stemming, we use the Turkish morphological analyzing tool Zemberek [4] and the Porter stemmer with NLTK [6]. Zemberek is commonly used in Turkish language studies [3, 10, 15, 16, 25]. It is a publicly available open-source program and contains java libraries that can be embedded in an application code. Some words are left without any changes if Zemberek cannot find their roots.

We use Python with NLTK [6] in our experiments. We extract features by using the eight pre-processing combinations, and then we use the term frequencies of all features in each review document to construct the feature vectors. We use only unigram features, as they tend to have better performance than n-grams for sentiment analysis [2]. For the feature vectors, we use the bag-of-words representation which is commonly used for text classification.

## 3.3. Classifiers

We use three machine-learning algorithms for text classification (NBM, SVM, and C4.5) since they have different characteristics and have been used widely for sentiment analysis and text classification [2, 12, 18, 23, 27]. More specifically, we use the NBM classifier for the Naïve Bayes Multinomial, Sequential Minimal Optimization (SMO) classifier for SVM, and J48 classifier for C4.5 in the Weka Data Mining Toolkit [26]. A linear kernel is set with the SMO algorithm, as we have high-dimensional feature space. We conduct five-fold cross validation for all of our experiments.

### 3.4. Performance Evaluation

Precision and Recall are two basic performance evaluation measures for text classification. Precision ($P$) is the percentage of correctly classified documents over all documents with respect to a particular class. Recall ($R$) is the percentage of the number of correctly classified documents over the total number of documents in a given class. The $F$-measure is defined as the harmonic mean of precision and recall [26] and is a widely used composite measure for text classification (which is given in Equation 1).

$$F = 2 * \frac{P * R}{P + R} \tag{1}$$

### 3.5. Feature Selection

We also apply our feature-selection method, which was developed especially for sentiment analysis [24]. In this method, we rank the features according to their probabilities in the positive and negative instances of the review datasets. The ranking score for feature $f$ is computed as in Equation 2:

$$score_f = \frac{p_f + q_f}{|p_f - q_f|}, \tag{2}$$

where $p_f$ and $q_f$ are the probabilities of feature $f$ in the positive and negative instances, which are computed as in Equations 3 and 4, respectively.

$$p_f = \frac{df_+^f + 0.5}{n^+ + 1.0} \tag{3}$$

$$q_f = \frac{df_-^f + 0.5}{n^- + 0.5} \tag{4}$$

In the above equations, $df_+^f$ and $df_-^f$ denote the document frequencies of feature $f$ in the positive and negative instances, respectively, and $n^+$ and $n^-$ are the numbers of instances in the positive and negative classes, respectively. We add small constants to the numerators and denominators in Equations 3 and 4 to avoid zero probabilities (as it is done in Lidstone smoothing). For each extracted feature in a dataset, we compute its ranking score according to Equation 2 and select the top n features to eliminate noisy or irrelevant features before they are used to classify the review documents.

## 4. Experiments and Results

In this section, we show the effects of the different pre-processing and classification methods so we can determine the best combinations for the Turkish and English review documents. We also apply feature selection to improve the accuracy of the classification and list the best stop words and punctuation marks for the sentiment analysis of review documents.

## 4.1. Performance of data pre-processing methods

After tokenizing the text into words (along with case normalization), we consider whether to keep punctuation marks (PN=yes/no), remove stop words (SR=yes/no), and perform stemming (ST=yes/no), giving us a total of eight combinations for the data pre-processing. After we apply the eight pre-processing combinations to each of the datasets, the numbers of the features extracted are listed in Table 6.

**Table 6**

Numbers of features extracted for each combination of data pre-processing methods for Turkish (TR) and English (EN) datasets

| | Preprocessing Methods | | | Movie #of features | | Book #of features | | DVD #of features | | Electronics #of features | | Kitchen #of features | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | PN | SR | ST | TR | EN | TR | EN | TR | EN | TR | EN | TR | EN |
| 1 | no | yes | no | 18,339 | 38,418 | 10,298 | 17,895 | 11,131 | 17,272 | 10,711 | 8622 | 9247 | 7732 |
| 2 | yes | yes | no | 18,352 | 38,424 | 10,309 | 17,906 | 11,140 | 17,283 | 10,721 | 8634 | 9258 | 7742 |
| 3 | no | yes | yes | 6078 | 24,961 | 3577 | 120,125 | 4161 | 12,196 | 3664 | 5888 | 3223 | 5242 |
| 4 | yes | yes | yes | 6091 | 24,967 | 3588 | 12,023 | 4170 | 12,207 | 3674 | 5900 | 3234 | 5252 |
| 5 | no | no | no | 18,565 | 38,863 | 10,500 | 18,295 | 11,334 | 17,663 | 10,901 | 8998 | 9436 | 8066 |
| 6 | yes | no | no | 18,578 | 38,869 | 10,511 | 18,306 | 11,343 | 17,676 | 10,911 | 9010 | 9447 | 8076 |
| 7 | no | no | yes | 6236 | 25,354 | 3718 | 12,376 | 4303 | 12,542 | 3803 | 6242 | 3361 | 5547 |
| 8 | yes | no | yes | 6249 | 25,360 | 3729 | 12,387 | 4312 | 12,553 | 3813 | 6254 | 3372 | 5557 |

We test all eight combinations of the data pre-processing on the three chosen classifiers (NBM, SVM, and J48). The classification results for the Turkish review datasets are presented in Tables 7 through 9. For each dataset and classifier, the best $F$-measure value is marked in **bold**.

**Table 7**

Classification results of data pre-processing methods for Turkish review datasets using NBM classifier

| # | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| 1 | 0.8208 | 0.8044 | 0.7799 | 0.8117 | 0.7650 |
| 2 | 0.8213 | 0.8089 | 0.7828 | 0.8103 | 0.7683 |
| 3 | 0.7522 | 0.7665 | 0.7227 | 0.7974 | 0.7558 |
| 4 | 0.7593 | 0.7708 | 0.7271 | 0.7933 | 0.7607 |
| 5 | **0.8258** | **0.8324** | 0.7928 | 0.8147 | **0.7767** |
| 6 | 0.8248 | 0.8317 | **0.7957** | **0.8155** | 0.7762 |
| 7 | 0.7588 | 0.8085 | 0.7474 | 0.8086 | 0.7637 |
| 8 | 0.7652 | 0.8107 | 0.7468 | 0.8071 | 0.7640 |

**Table 8**
Classification results of data pre-processing methods for Turkish review datasets using SVM classifier

| # | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| 1 | 0.8043 | 0.7760 | 0.7138 | 0.7678 | 0.7410 |
| 2 | 0.8039 | 0.7798 | 0.7116 | 0.7671 | 0.7450 |
| 3 | 0.7429 | 0.7190 | 0.6721 | 0.7382 | 0.7099 |
| 4 | 0.7532 | 0.7206 | 0.6657 | 0.7359 | 0.7128 |
| 5 | 0.8132 | **0.7960** | **0.7327** | **0.7764** | 0.7364 |
| 6 | **0.8161** | 0.7955 | 0.7320 | 0.7707 | **0.7407** |
| 7 | 0.7497 | 0.7630 | 0.7113 | 0.7409 | 0.7227 |
| 8 | 0.7533 | 0.7639 | 0.7085 | 0.7444 | 0.7170 |

**Table 9**
Classification results of data pre-processing methods for Turkish review datasets using J48 classifier

| # | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| 1 | 0.6850 | 0.7020 | 0.6684 | 0.7060 | 0.6710 |
| 2 | 0.7097 | 0.6966 | 0.6642 | 0.7021 | 0.6733 |
| 3 | 0.6753 | 0.6910 | 0.6407 | 0.6996 | 0.6577 |
| 4 | 0.6913 | 0.6820 | 0.6428 | 0.6956 | 0.6607 |
| 5 | **0.7018** | **0.7020** | **0.6929** | 0.7221 | **0.6691** |
| 6 | 0.6954 | 0.7019 | 0.6886 | **0.7371** | 0.6647 |
| 7 | 0.6771 | 0.7010 | 0.6857 | 0.7128 | 0.6757 |
| 8 | 0.6861 | 0.6979 | 0.6914 | 0.7107 | 0.6678 |

As can be seen in Tables 7 through 9, the best results are all obtained with the NBM classifier for each dataset. The best pre-processing combination may change for each classifier and dataset; however, the best pre-processing combinations are 5 and 6 for all of the datasets and classifiers. Among the results presented in Tables 7 through 9, it can be observed that Combination 6 with NBM produces the best results for the DVD and electronic review datasets; however, for the movie, book, and kitchen review datasets, it is Combination 5 and NBM that give the best results. Combination 6 is the most conservative since it essentially retains all of the features in a dataset by keeping all punctuation patterns and stop words and not performing stemming. Combination 5 differs only in removing all of the punctuation patterns. The most aggressive method is Combination 3, where removing the punctuation patterns and stop words and performing stemming actually shows the worst performance for most of the datasets.

By examining the punctuation patterns in Table 5, we can see that some patterns such as happy and sad faces, exclamations, and question marks may carry meanings for sentiments, but other patterns such as periods, commas, colons, and semicolons do not seem to express sentiments. Accordingly, we also try to keep some punctuation patterns while removing others, creating three more combinations (where $6^1$ keeps Patterns 1 through 4 and 6 through 7 in Table 5, $6^2$ keeps all patterns in $6^1$ plus Patterns 10 through 11, and $6^3$ keeps all patterns in $6^2$ plus Pattern 5). The classification results for these combinations along with those for Combinations 5 and 6 are given in Table 10. As can be seen, the performance remains mixed: for some datasets, the new combinations are helpful, but Combination 6 is still the best for others. To see the impact of these pre-processing methods, we conduct a univariate ANOVA analysis for all of the combinations in Table 6; these results show that there are no significant differences among these methods at a 95% confidence level. This leads us to conclude that we should generally treat punctuation patterns as regular features and let the feature-selection methods do further cutting if needed, since different sets of punctuation patterns are helpful for different datasets.

**Table 10**

Classification results of additional data pre-processing methods for Turkish review datasets using NBM classifier

| # | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| $6^1$ | **0.8269** | **0.8324** | 0.7928 | 0.8148 | **0.7767** |
| $6^2$ | 0.8259 | **0.8324** | 0.7928 | 0.8148 | **0.7767** |
| $6^3$ | 0.8244 | 0.8302 | **0.7943** | **0.8155** | 0.7759 |

Also observed in Tables 7 to 9 is the fact that Combinations 5 and 6 perform better than the other combinations. This indicates that keeping the stop words and not performing stemming are desirable for the sentiment analysis of Turkish reviews. Stop words are typically identified manually, and some lists are longer than others. We create our Turkish stop word list according to our datasets. Similar to the case for the punctuation patterns, our results indicate that keeping the stop words helps us achieve better performance no matter if the stop word list is long or short. A detailed analysis shows that some common stop words can actually express sentiments such as "what" in "what a performance," "too" in "too small," and "not" in "not quite interesting." In addition, the decrease in the performance of Combinations 6 through 8 suggests that stemming can hurt the classification performance as well. This is possibly due to the agglutinative property of the Turkish language. Compared with other works, Kaya et al. [16] and Sevindi [25] only consider some pre-processing tasks and show that there is a decrease in performance when stemming and stop word removal are used. Similarly, Erogul [15] shows that stemming makes no significant improvement. The results presented in our study also confirm these results.

For the English review datasets, a similar process was followed; the results among all classifiers and across all combinations for the data pre-processing are included in Tables 11 through 13. One thing that is different from the Turkish datasets is that, for the English review datasets, SVM produces the best results for the movie, electronics, and kitchen review datasets, while NBM is the best performer for the DVD and book review datasets. Across all combinations, however, the best results fluctuate between Combinations 5 and 6, indicating that we should also keep punctuation patterns and stop words and not perform stemming for the English review datasets. This confirms the observations made in Pang et al. [23]. Also different from the Turkish reviews, the statistical analysis indicates that there are significant differences among the pre-processing methods for the English reviews at a 95% confidence level. However, the post-hoc analysis shows that there is no significant difference between Combinations 5 and 6.

**Table 11**

Classification results of data pre-processing methods for English review datasets using NBM classifier

| # | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| 1 | 0.8064 | 0.7441 | 0.7845 | 0.7628 | 0.7928 |
| 2 | 0.8039 | 0.7497 | 0.7796 | 0.7643 | 0.7871 |
| 3 | 0.7650 | 0.6787 | 0.7328 | 0.6957 | 0.7171 |
| 4 | 0.7620 | 0.6795 | 0.7343 | 0.6949 | 0.7207 |
| 5 | **0.8199** | 0.7597 | **0.7879** | **0.7657** | 0.8014 |
| 6 | 0.8129 | **0.7619** | 0.7836 | 0.7629 | **0.8099** |
| 7 | 0.7800 | 0.6915 | 0.7514 | 0.7193 | 0.7557 |
| 8 | 0.7740 | 0.6996 | 0.7521 | 0.7143 | 0.7521 |

**Table 12**

Classification results of data pre-processing methods for English review datasets using SVM classifier

| # | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| 1 | 0.8364 | 0.7393 | 0.7548 | 0.7613 | 0.8021 |
| 2 | 0.8374 | 0.7421 | 0.7534 | 0.7591 | 0.8014 |
| 3 | 0.7575 | 0.6745 | 0.6946 | 0.6991 | 0.7293 |
| 4 | 0.7560 | 0.6791 | 0.6946 | 0.6948 | 0.7285 |
| 5 | **0.8485** | **0.7499** | **0.7663** | 0.7834 | **0.8157** |
| 6 | 0.8480 | 0.7485 | 0.7649 | **0.7856** | 0.8136 |
| 7 | 0.7785 | 0.7096 | 0.7219 | 0.7477 | 0.7621 |
| 8 | 0.7775 | 0.7097 | 0.7262 | 0.7399 | 0.7671 |

**Table 13**
Classification results of data pre-processing methods for English review datasets using J48 classifier

| # | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| 1 | 0.6818 | 0.6298 | 0.6627 | 0.6719 | 0.7071 |
| 2 | 0.6754 | 0.6343 | 0.6717 | 0.6726 | 0.7070 |
| 3 | 0.6419 | 0.5925 | 0.6326 | 0.6213 | 0.6493 |
| 4 | 0.6410 | 0.6036 | 0.6255 | 0.6276 | 0.6486 |
| 5 | **0.6784** | 0.6393 | 0.6789 | **0.6828** | **0.7107** |
| 6 | 0.6769 | **0.6407** | **0.6821** | 0.6750 | 0.7093 |
| 7 | 0.6494 | 0.6135 | 0.6186 | 0.6198 | 0.6571 |
| 8 | 0.6499 | 0.5948 | 0.6149 | 0.6212 | 0.6449 |

To summarize, we see some similarities between the Turkish and English reviews in that we should keep the punctual patterns and stop words and not perform stemming for the data pre-processing, leading us to use the same setting as the baselines for further study. In addition, NBM seems to be the most suitable classifier for sentiment analysis since sentiment-expressing words tend to have low frequencies within a document yet relatively high frequencies across different documents.

Moreover, SVM can also perform well for some English review datasets, while NBM looks like the dominant classifier for the Turkish reviews. Finally, the performance results for the English reviews are generally higher than those for the Turkish reviews, possibly related to the differences between the two languages in terms of vocabularies, writing styles, and the agglutinative property of the Turkish language.

## 4.2. Performance of feature selection

We use pre-processing Combination 6, which does not apply stemming yet does include stop words and punctuation marks; then, we apply our feature-selection method on both the Turkish and English review datasets to see which stop words and punctuation marks are selected. As our method computes a ranking score for each feature, we select six feature sizes of 500, 1000, 1500, 2000, 2500, and 3000, since we also observed in our previous studies that feature sizes up to 3000 tend to give good classification performance. We pick the top-ranked features of a desirable size $n$ based on the computed scores, and they are run against three classifiers (NBM, SVM, and J48) for each review dataset. As NBM is the best classifier for the Turkish datasets and is also the best classifier in most of the cases of the English datasets, we present only the results for NBM in Tables 14 and 15 to save space.

**Table 14**
Classification results of feature-selection method for Turkish datasets using NBM classifier

| | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| 500 | 0.8502 | 0.8996 | 0.8735 | 0.8658 | 0.8518 |
| 1000 | 0.8757 | 0.9113 | 0.8982 | 0.8704 | **0.8790** |
| 1500 | 0.8944 | **0.9150** | **0.9136** | **0.8996** | 0.8765 |
| 2000 | 0.9046 | 0.8992 | 0.9050 | 0.8968 | 0.8721 |
| 2500 | 0.9082 | 0.8913 | 0.8956 | 0.8894 | 0.8595 |
| 3000 | **0.9112** | 0.8776 | 0.8819 | 0.8806 | 0.8365 |
| All features | 0.8248 | 0.8317 | 0.7957 | 0.8155 | 0.7762 |

**Table 15**
Classification results of feature-selection method for English datasets using NBM classifier

| | Movie Reviews | Book Reviews | DVD Reviews | Electronics Reviews | Kitchen Reviews |
|---|---|---|---|---|---|
| 500 | 0.6980 | 0.7952 | 0.8191 | 0.8223 | 0.8580 |
| 1000 | 0.8219 | 0.8552 | 0.8773 | 0.8718 | 0.9090 |
| 1500 | 0.8967 | 0.8902 | 0.8845 | 0.8724 | 0.9084 |
| 2000 | 0.9134 | 0.8873 | 0.9111 | **0.8878** | **0.9106** |
| 2500 | **0.9410** | 0.8901 | **0.9169** | 0.8877 | 0.9099 |
| 3000 | 0.9355 | **0.9162** | 0.9155 | 0.8812 | 0.9026 |
| All features | 0.8129 | 0.7619 | 0.7836 | 0.7629 | 0.8099 |



As can be easily seen in Tables 14 and 15, applying feature selection improves the classification accuracy of NBM by removing irrelevant or noisy features (as observed in [21]). Similar trends can also be observed for the SVM and J48 classifiers; however, these results are not presented here to save space.

In Table 16, we list the numbers of stop words and punctuation marks chosen by the QER feature-selection method for the best feature sizes of the Turkish and English review datasets. As shown in Table 16, only one or (at most) two punctuation patterns and fewer than 10% of the stop words are selected by the feature-selection algorithm. Therefore, we can conclude that not all stop words and punctuation patterns are helpful.

**Table 16**
Number of stop words (SW) and punctuation (PN) patterns chosen for datasets

| | #of features selected | #of SW selected | Total #of SW | #of PN patterns selected | Total #of PN patterns |
|---|---|---|---|---|---|
| Movies(TR) | 3000 | 24 | 248 | 1 | 13 |
| Movies(EN) | 2500 | 11 | 474 | 0 | 6 |
| Books(TR) | 1500 | 49 | 219 | 0 | 11 |
| Books(EN) | 3000 | 26 | 414 | 1 | 11 |
| DVDs(TR) | 1500 | 29 | 228 | 0 | 9 |
| DVDs(EN) | 2500 | 21 | 409 | 1 | 11 |
| Electronics(TR) | 1500 | 19 | 213 | 1 | 10 |
| Electronics(EN) | 2000 | 45 | 399 | 1 | 12 |
| Kitchen(TR) | 1000 | 13 | 205 | 2 | 11 |
| Kitchen(EN) | 2000 | 35 | 351 | 0 | 10 |

In Table 17, we list the most selected stop words and punctuation patterns for all of the Turkish and English datasets, which may be helpful to other researchers in their future studies. As can be seen in Table 17, the most valuable stop words and punctuation patterns are different for the different languages, which shows that language-specific pre-processing methods should be applied for sentiment analysis.

**Table 17**
Selected stop words and punctuation patterns for English and Turkish Electronics datasets

| | Turkish | English |
|---|---|---|
| Stop words | asla, birkez, birşeyi bize, diğeri, elbette, gene, herkesin, karşın, kendisi, kimisi, mi, mu, neden, nedir, on, onlara, onlarda, tamam | (mr, mostly, further, ours, former, click, seven, whoever, latter, beforehand, therefore, thirty, wherever, ten, caption, moreover, overall, onto, sixty, hence, thru, besides, thereafter, million |
| Punctuation marks | ? | =) =)) =))) |

## 5. Conclusions

In this paper, we examined the impacts of data pre-processing on the sentiment analysis of Turkish and English reviews and showed the similarities and differences between an agglutinative language and English. For data pre-processing, we tried different combinations of punctuation patterns, stop words, and stemming. All of these methods are tested against ten datasets of Turkish and English reviews, using common text classifiers that included Naïve Bayes Multinomial (NBM), Support Vector Machines (SVM), and Decision Trees (J48). Our results show that, for data pre-processing, it is

important to keep punctuation patterns and stop words as features and not perform stemming for both Turkish and English reviews. This is because certain punctuation patterns, stop words, and non-stemmed words can carry sentiment meanings in some datasets, and removing them can negatively affect the performance of sentiment analysis. Accordingly, we decide to keep all of the original words as features and leave it to feature selection to cut down the number of irrelevant features for possible improvements in classification. For all of the Turkish review datasets and some of the English review datasets in our experiments, the best results were all obtained with the NBM classifier. SVM performs best for some English review datasets, while NBM looks like the dominant classifier for the Turkish reviews.

We also observe that the performance results for English reviews are generally higher than those for Turkish reviews, possibly related to the differences between the two languages in terms of vocabularies, writing styles, and the agglutinative property of the Turkish language.

Recently, deep-learning models have been applied to sentiment analysis with significantly improved results [17]. In particular, the words are represented with real-numbered vectors so that related words will have closer distance values, allowing us to model many kinds of word relationships such as morphology, agglutinations, punctuations, and even meanings [8]. How such relationships can be explicitly distinguished and applied to feature selection would be interesting to explore, as the results may help us improve the performance of classical machine learning methods and achieve better comparisons with deep-learning models.

## Acknowledgements

*The research presented in this paper was supported by Çukurova University Academic Research Project Unit (under Grant No. FDK-2015-3833) and the Scientific and Technological Research Council of Turkey (Scholarship TUBITAK 2214-A).*

## References

[1] Abbasi A., Chen H., Salem A.: Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. In: *ACM Transactions on Information Systems ...*, vol. 26(3), pp. 1–34, 2008. ISSN 10468188. `http://dx.doi.org/10.1145/1361684.1361685`.

[2] Agarwal B., Mittal N.: Prominent feature extraction for review analysis: an empirical study. In: *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28(3), pp. 485–498, 2016. ISSN 0952-813X. `http://dx.doi.org/10.1080/0952813X.2014.977830`.

[3] Akba F., Uçan A., Sezer E., Sever H.: Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews. In: *8th European Conference on Data Mining*, pp. 180–184. Lisbon, Portugal, 2014. ISBN 9789898704108. `http://humir.cs.hacettepe.edu.tr/file/AkbaFUcanA.pdf`.

[4] Akın A.A., Akın M.D.: Zemberek, An Open Source Nlp Framework for Turkic Languages. In: *Structure*, vol. 10, pp. 1–5, 2007. http://zemberek.googlecode.com/files/zemberek_makale.pdf.

[5] Asgarian E., Kahani M., Sharifi S.: The Impact of Sentiment Features on the Sentiment Polarity Classification in Persian Reviews. In: *Cognitive Computation*, vol. 10(1), pp. 117–135, 2018. ISSN 1866-9956. http://dx.doi.org/10.1007/s12559-017-9513-1.

[6] Bird S., Klein E., Loper E.: *Natural Language Processing with Python*. O'Reilly, 2009. http://www.nltk.org/book_1ed/.

[7] Blitzer J., Dredze M., Pereira F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: *45th Annual Meeting-Association for Computational Linguistics*, pp. 440–447. 2007. ISBN 9781424491131. ISSN 0736587X. http://dx.doi.org/10.1109/IRPS.2011.5784441.

[8] Bojanowski P., Grave E., Joulin A., Mikolov T.: Enriching Word Vectors with Subword Information. In: *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[9] Çakici R.: *Wide-coverage parsing for Turkish*. Ph.D. thesis, PhD Thesis, University of Edinburgh, 2009. http://hdl.handle.net/1842/3807.

[10] Cetin M., Amasyali M.F.: Supervised and traditional term weighting methods for sentiment analysis. In: *21st Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, 2013. ISSN 0162-8828. http://dx.doi.org/10.1109/SIU.2013.6531173.

[11] Demirtas E., Pechenizkiy M.: Cross-lingual polarity detection with machine translation. In: *Second International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '13*, pp. 1–8. ACM Press, New York, New York, USA, 2013. ISBN 9781450323321. http://dx.doi.org/10.1145/2502069.2502078.

[12] Despotovic V., Tanikic D.: Sentiment Analysis of Microblogs Using Multilayer Feed-Forward Artificial Neural Networks. In: *COMPUTING AND INFORMATICS*, vol. 36(5), pp. 1127–1142, 2017. ISSN 2585-8807. http://www.cai.sk/ojs/index.php/cai/article/viewArticle/2017_5_1127.

[13] Devitt A., Ahmad K.: Sentiment polarity identification in financial news: a cohesion-based approach. In: *Proceedings of Annual Meeting of the Association of Computational Linguistics*, (June), pp. 984–991, 2007. ISSN 0736587X. http://dx.doi.org/10.1.1.143.7157.

[14] Duwairi R., El-Orfali M.: A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. In: *Journal of Information Science*, vol. 40(4), pp. 501–513, 2014. ISSN 0165-5515. http://dx.doi.org/10.1177/0165551514534143.

[15] Eroğul U.: *Sentiment Analysis in Turkish.* Ph.D. thesis, 2009. http://dx.doi.org/10.1007/s13398-014-0173-7.2.

[16] Kaya M., Fidan G., Toroslu I.H.: Sentiment Analysis of Turkish Political News. In: *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pp. 174–180. IEEE, Macau, China, 2012. ISBN 978-1-4673-6057-9. http://dx.doi.org/10.1109/WI-IAT.2012.115.

[17] Kim Y.: Convolutional Neural Networks for Sentence Classification. In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751. 2014. http://nlp.stanford.edu/sentiment/http://arxiv.org/abs/1408.5882.

[18] Liu Y., Bi J.W., Fan Z.P.: Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. In: *Expert Systems with Applications*, vol. 80, pp. 323–339, 2017. ISSN 09574174. http://dx.doi.org/10.1016/j.eswa.2017.03.042.

[19] Mladenović M., Mitrović J., Krstev C., Vitas D.: Hybrid sentiment analysis framework for a morphologically rich language. In: *Journal of Intelligent Information Systems*, vol. 46(3), pp. 599–620, 2016. ISSN 0925-9902. http://dx.doi.org/10.1007/s10844-015-0372-5.

[20] Nicholls C., Song F.: Comparison of Feature Selection Methods for Sentiment Analysis. In: *Advances in Artificial Intelligence*, pp. 286–289. Springer, Berlin, Heidelberg, 2010. ISBN 978-3-642-13059-5. http://dx.doi.org/10.1007/978-3-642-13059-5{_}30.

[21] Pang B., Lee L.: A sentimental education. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, pp. 271–es. Association for Computational Linguistics, Morristown, NJ, USA, 2004. ISSN 1554-0669. http://dx.doi.org/10.3115/1218955.1218990.

[22] Pang B., Lee L.: Opinion Mining and Sentiment Analysis. In: *Foundations and Trends® in Information Retrieval*, vol. 2(1–2), pp. 1–135, 2008. ISSN 1554-0669. http://dx.doi.org/10.1561/1500000011.

[23] Pang B., Lee L., Vaithyanathan S.: Thumbs up? In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, vol. 10, pp. 79–86. Association for Computational Linguistics, Morristown, NJ, USA, 2002. http://dx.doi.org/10.3115/1118693.1118704.

[24] Parlar T., Özel S., Song F.: QER: a new feature selection method for sentiment analysis. In: *Human-centric Computing and Information Sciences*, vol. 8(1), p. 10, 2018. ISSN 21921962. http://dx.doi.org/10.1186/s13673-018-0135-8.

[25] Sevindi B.I.: *Türkçe Metinlerde Denetimli ve Sözlük Tabanlı Duygu Analizi Yaklaşımlarının Karşılaştırılması.* Ph.D. thesis, MSc Thesis, Gazi University, 2013.

[26] Witten I.H., Frank E., Hall M.A.: *Data mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2011.

[27] Yang D.H., Yu G.: A method of feature selection and sentiment similarity for Chinese micro-blogs. In: *Journal of Information Science*, vol. 39(4), pp. 429–441, 2013. ISSN 0165-5515. http://dx.doi.org/10.1177/0165551513480308.

[28] Zheng L., Wang H., Gao S.: Sentimental feature selection for sentiment analysis of Chinese online reviews. In: *International Journal of Machine Learning and Cybernetics*, vol. 9(1), pp. 75–84, 2018. ISSN 1868-8071. http://dx.doi.org/10.1007/s13042-015-0347-4.

## Affiliations

**Tuba Parlar**
Mustafa Kemal University, Hatay, Turkiye, tparlar@mku.edu.tr,
ORCID ID: https://orcid.org/0000-0002-8004-6150

**Selma Ayşe Özel**
Çukurova University, Adana, Turkiye, saozel@cu.edu.tr,
ORCID ID: https://orcid.org/0000-0001-9201-6349

**Fei Song**
University of Guelph, Ontario, Canada, fsong@uoguelph.ca,
ORCID ID: https://orcid.org/0000-0003-3036-9696

**Received**: 20.10.2018
**Revised**: 13.03.2019
**Accepted**: 13.03.2019