**Analysis of data pre-processing methods for the sentiment analysis of reviews**

1. Introduction

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. 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.

2. Related Work

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.

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.

3. Materials and Methods

3.1. Data sets

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.

3.3 Classifiers

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.

3.5. Feature Selection

4. Experiments and Results

4.1. Performance of data pre-processing methods

4.1. Performance of feature selection

5. Conclusions