SETSCI Conference Indexing System,
Volume 2 (2018), 223-223

ISAS 2018
I[st] International Symposium on Innovative Approaches in Scientific Studies
April 11-13, 2018, Kemer-Antalya, Turkey

Comparison of Text Classification Performances in Balanced and Imbalanced Datasets

Mehmet Fatih KARACA[1*], Şafak BAYIR[2]
[1] Department of Computer Technologies, Gaziosmanpaşa University, TURKEY
[2] Department of Computer Engineering, Karabük University, Karabuk, TURKEY
[*]Corresponding author: mfkaraca@gmail.com
[+]Speaker: mfkaraca@gmail.com
Presentation/Paper Type: Oral / Abstract

Abstract- Text mining, which aims to derive unknown and useful information from available textual data, is one of the subfield of data mining. Text mining transforms unstructured textual data into structured form by utilizing various methods. Data mining techniques can also be applied to textual data as a result of transformation to structured form. Text classification which is one of the widely favoured subject of text mining is the process of assigning texts into predefined classes. As a result of this, classification is realized more rapidly and consistently.

In this study, text classification performances in balanced and imbalanced datasets were compared. Corpora which consist of Turkish and English texts were utilized. The features of 4 datasets including 5 classes in each were as follows: Corpus 1 and Corpus 3 include Turkish contents and Corpus 2 and Corpus 4 include English contents. 3375 training and 1125 test documents were included in the Corpus 1 and Corpus 2 which are balanced datasets, whereas 1825 training and 825 test documents were included in the Corpus 3 and Corpus 4 which are imbalanced datasets.

Documents included in datasets were pre-processed, weighted as binary and tf-idf and document vectors were obtained. kNN was preferred for text classification and Manhattan Distance, Harmonic Mean, Inner Product, Squared Chord and Dice's Coefficient were selected for the measurement of similarities of document vectors.

Results were taken into consideration in terms of classification success as well as process time. Since the number of the documents within the classes are equal, it is seen that more successful classification was obtained in terms of average values in both Turkish and English content balanced datasets but the process time was longer.

Keywords- text mining, text classification, kNN, balanced dataset, imbalanced dataset.