عنوان مقاله:

An Effective Method of Feature Selection in Persian Text for Improving the Accuracy of Detecting Request in Persian Messages on Telegram

محل انتشار:

فصلنامه سیستم های اطلاعاتی و مخابرات, دوره 8, شماره 4 (سال: 1399)

تعداد صفحات اصل مقاله: 14

نویسندگان:

Zahra Khalifeh Zadeh - *Faculty of Computer Engineering, Yazd University, Iran*

Mohammad Ali Chahooki - *Faculty of Computer Engineering, Yazd University, Iran*

خلاصه مقاله:

In recent years, data received from social media has increased exponentially. They have become valuable sources of information for many analysts and businesses to expand their business. Automatic document classification is an essential step in extracting knowledge from these sources of information. In automatic text classification, words are assessed as a set of features. Selecting useful features from each text reduces the size of the feature vector and improves classification performance. Many algorithms have been applied for the automatic classification of text. Although all the methods proposed for other languages are applicable and comparable, studies on classification and feature selection in the Persian text have not been sufficiently carried out. The present research is conducted in Persian, and the introduction of a Persian dataset is a part of its innovation. In the present article, an innovative approach is presented to improve the performance of Persian text classification. The authors extracted ۸۵,۰۰۰ Persian messages from the Idekav-system, which is a Telegram search engine. The new idea presented in this paper to process and classify this textual data is on the basis of the feature vector expansion by adding some selective features using the most extensively used feature selection methods based on Local and Global filters. The new feature vector is then filtered by applying the secondary feature selection. The secondary feature selection phase selects more appropriate features among those added from the first step to enhance the effect of applying wrapper methods on classification performance. In the third step, the combined filter-based methods and the combination of the results of different learning algorithms have been used to achieve higher accuracy. At the end of the three selection stages, a method was proposed that increased accuracy up to ۰.۹۴۵ and reduced training time and calculations in the Persian dataset.

کلمات کلیدی:

Feature Selection; Text Mining; Classification Accuracy; Machine Learning; Ensemble Classifier