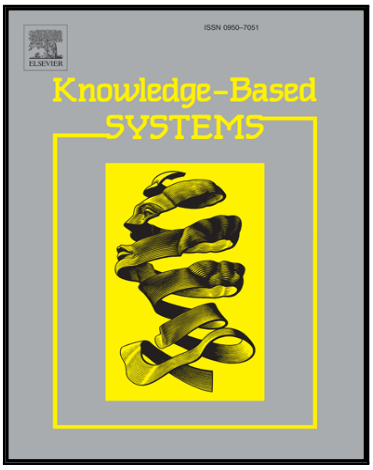

An Automated Text Categorization Framework based on Hyperparameter Optimization

Eric S. Tellez, Daniela Moctezuma, Sabino Miranda-Jiménez, Mario Graff

PII: S0950-7051(18)30121-7
DOI: 10.1016/j.knosys.2018.03.003
Reference: KNOSYS 4252

To appear in: *Knowledge-Based Systems*

Received date: 20 September 2017
Revised date: 10 January 2018
Accepted date: 1 March 2018

An Automated Text Categorization Framework based on Hyperparameter Optimization

Eric S. Tellez[a,c], Daniela Moctezuma[a,b,*], Sabino Miranda-Jiménez[a,c], Mario Graff[a,c]

[a] *CONACyT Consejo Nacional de Ciencia y Tecnología, Dirección de Cátedras, Insurgentes Sur 1582, Crédito Constructor 03940, Ciudad de México, México.*
[b] *Centro de Investigación en Ciencias de Información Geoespacial, Circuito Tecnopolo Norte No. 117, Col. Tecnopolo Pocitos II, C.P. 20313, Aguascalientes, Ags, México.*
[c] *INFOTEC Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Circuito Tecnopolo Sur No 112, Fracc. Tecnopolo Pocitos II, Aguascalientes 20313, México.*

Abstract

A great variety of text tasks such as topic or spam identification, user profiling, and sentiment analysis can be posed as a supervised learning problem and tackled using a text classifier. A text classifier consists of several subprocesses, some of them are general enough to be applied to any supervised learning problem, whereas others are specifically designed to tackle a particular task using complex and computational expensive processes such as lemmatization, syntactic analysis, etc. Contrary to traditional approaches, we propose a minimalist and multi-propose text-classifier able to tackle tasks independently of domain and language. We named our approach μTC. Our approach is composed of several easy-to-implement text transformations, text representations, and a supervised learning algorithm. These pieces produce a competitive classifier in several challenging domains such as informally written text. We provide a detailed description of μTC along with an extensive experimental comparison with relevant state-of-the-art methods, i.e., μTC was compared on 30 different datasets. Regarding accuracy, μTC obtained the best performance in 20 datasets while achieves competitive results in the remaining ones. The compared datasets include several problems like topic and polarity classification, spam detection, user profiling and authorship attribution. Furthermore, our approach allows the usage of the technology even without an in-depth knowledge of machine learning and natural language processing.

Keywords: text classification, hyperparameter optimization, text modelling

[*] Corresponding author: email: dmoctezuma@centrogeo.edu.mx tel.: +52-449-9945150 ext. 5203

Email addresses: eric.tellez@infotec.mx (Eric S. Tellez), sabino.miranda@infotec.mx (Sabino Miranda-Jiménez), mario.graff@infotec.mx (Mario Graff)