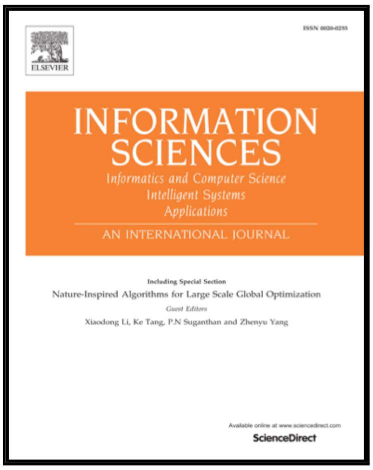
Accepted Manuscript



An Analysis of Hierarchical Text Classification Using Word Embeddings

Roger Alan Stein, Patrícia A. Jaques, João Francisco Valiati

PII: S0020-0255(18)30693-5
DOI: https://doi.org/10.1016/j.ins.2018.09.001
Reference: INS 13912

To appear in: *Information Sciences*

Received date: 17 February 2018
Revised date: 29 August 2018
Accepted date: 1 September 2018

Please cite this article as: Roger Alan Stein, Patrícia A. Jaques, João Francisco Valiati, An Analysis of Hierarchical Text Classification Using Word Embeddings, *Information Sciences* (2018), doi: https://doi.org/10.1016/j.ins.2018.09.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

An Analysis of Hierarchical Text Classification Using Word Embeddings

Roger Alan Stein[a], Patrícia A. Jaques[a], João Francisco Valiati[b]

[a]*Programa de Pós-Graduação em Computação Aplicada—PPGCA*
Universidade do Vale do Rio dos Sinos—UNISINOS
Av. Unisinos, 950, São Leopoldo, RS, Brazil
[b]*Artificial Intelligence Engineers—AIE*
Rua Vieira de Castro, 262, Porto Alegre, RS, Brazil

Abstract

Efficient distributed numerical word representation models (word embeddings) combined with modern machine learning algorithms have recently yielded considerable improvement on automatic document classification tasks. However, the effectiveness of such techniques has not been assessed for the hierarchical text classification (HTC) yet. This study investigates application of those models and algorithms on this specific problem by means of experimentation and analysis. We trained classification models with prominent machine learning algorithm implementations—fastText, XGBoost, SVM, and Keras' CNN—and noticeable word embeddings generation methods—GloVe, word2vec, and fastText—with publicly available data and evaluated them with measures specifically appropriate for the hierarchical context. FastText achieved an LCAF$_1$ of 0.893 on a single-labeled version of the RCV1 dataset. An analysis indicates that using word embeddings and its flavors is a very promising approach for HTC.

Keywords: Hierarchical Text Classification, Word Embeddings, Gradient Tree Boosting, fastText, Support Vector Machines

ACCEPTED MANUSCRIPT

1. Introduction

Electronic text processing systems are ubiquitous nowadays—from instant messaging applications in smartphones to virtual repositories with millions of documents—and have created some considerable challenges to address users new information needs. One of such endeavors is classifying automatically some of this textual data so that information system users can more easily retrieve, extract, and manipulate information to recognize patterns and generate knowledge. Organizing electronic documents into categories has become of increasing interest for many people and organizations [18, 27]. Text classification (TC)—a.k.a. text categorization, topic classification—is the field that studies solutions for this problem, and uses a combination of knowledge areas such as Information Retrieval, Artificial Intelligence, Natural Language Processing (NLP), Data Mining, Machine Learning, and Statistics. This is usually regarded as a supervised machine learning problem, where a model can be trained from several examples and then used to classify a previously unseen piece of text [37, 11].

TC tasks usually have two or a just few classes, for example, automatic email categorization, spam detection, customer request routing, etc. Classification tasks with a high number of possible target classes are studied as a further extension of the TC problem because they present some particular issues, which demand specific addressing or solutions. Many important real-world classification problems consist of a very large number of often very similar categories that are organized into a class hierarchy or taxonomy [27, 38]. This is where the hierarchical classification (HC) arises: it is a particular type of structured

URL: rogerstein@gmail.com (Roger Alan Stein), pjaques@unisinos.br (Patrícia A. Jaques), joao.valiati@ai-engineers.com (João Francisco Valiati)

Download English Version:

https://daneshyari.com/en/article/10145943

Download Persian Version:

https://daneshyari.com/article/10145943

Daneshyari.com