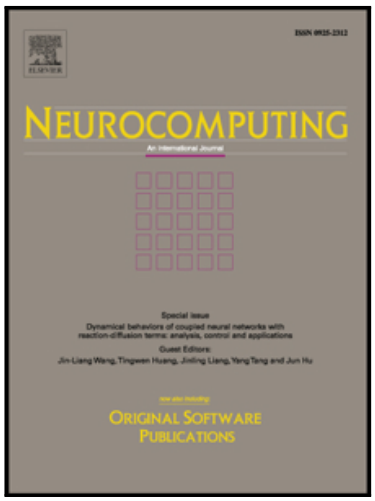


# An Evaluation Study on Text Categorization Using Automatically Generated Labelled Dataset

Dengya Zhu , Kok Wai Wong

PII: S0925-2312(17)30569-6
DOI: 10.1016/j.neucom.2016.04.072
Reference: NEUCOM 18282

To appear in: *Neurocomputing*

Received date: 7 October 2015
Revised date: 28 December 2015
Accepted date: 10 April 2016

Please cite this article as: Dengya Zhu , Kok Wai Wong , An Evaluation Study on Text Categorization Using Automatically Generated Labelled Dataset, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2016.04.072

# An Evaluation Study on Text Categorization Using Automatically Generated Labelled Dataset

Dengya Zhu[#1], Kok Wai Wong[#2]

[#1]*School of Information Systems, Curtin University, GPO Box U1987, Perth, Western Australia*

[#2]*School of Engineering and Information Technology, Murdoch University, South St, Perth, Western Australia, 6150*

[1]d.zhu@curtin.edu.au, [2]k.wong@murdoch.edu.au

**Abstract**

Naïve Bayes, k-Nearest Neighbours, Adaboost, Support Vector Machines and Neural Networks are five among others commonly used text classifiers. Evaluation of these classifiers involves a variety of factors to be considered including benchmark used, feature selections, parameter settings of algorithms, and the measurement criteria employed. Researchers have demonstrated that some algorithms outperform others on some corpus, however, inconsistency of human labelling and high dimensionality of feature spaces are two issues to be addressed in text categorization. This paper focuses on evaluating the five commonly used text classifiers by using an automatically generated text document collection which is labelled by a group of experts to alleviate subjectivity of human category assignments, and at the same time to examine the influence of the number of features on the performance of the algorithms.

Keywords

Text mining; Text categorization; Machine learning; Evaluation; Feature selection; benchmark collection

## 1. Introduction

Text categorization (a.k.a. classification) is defined as "the automated assignment of natural language texts to predefined categories based on their content" [1]. Let $D = \{\mathbf{d}_j \mid \mathbf{d}_j \in D, j = 1, \ldots, N\}$ is a text collection and for each document $\mathbf{d}_j \in D$, it has been assigned a unique category $c_i$ from a limited set of categories (or labels) $C = \{c_i \mid c_i \in C, i = 1, \ldots, M\}$[1]. Using this labelled dataset as training data, a classification model will be trained. For a given test instance (or example) for which the class label is unknown, the trained model will predict a label for the instance. Text categorization is a kind of supervised learning and has been widely applied in the areas such as language identification, information retrieval, opinion mining, spam filtering, and email routing [2]. With the recent explosion of information on the Web, text categorization is becoming increasingly important as an approach to managing and organizing the huge volume of information on the Web. Many algorithms such as Boostexter [3] and Support Vector Machines (SVMs) [4] have been developed and introduced for this purpose. Consequently, the evaluation of the effectiveness of the algorithms is playing an important role for both researchers and practitioners.

To evaluate a text categorization algorithm, the first element to be considered is the training document collection to be used. Many such document collections have been developed for evaluation purposes. The widely used benchmark collections for text

[1] This paper considers only this hard version of classification problem. Soft version classification allows a document to be assigned any number of labels.