IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

"Text classification using Bag of Words Representation and classifier fusion using Matlab"

KAVYA S.N[1]
[1]Assistant Professor, Department of Computer Science,
MMK & SDM MMV College, Mysore-570004.

BHARGAVI H.G[2]
[2]Assistant Professor, Department of Computer Science,
MMK & SDM MMV College, Mysore-570004.

Abstract: *A text classification plays a major role in text processing and it is an important instance of the classification problem with unique challenges and requirements. We introduce a single-label supervised text classification or multi-label supervised text classification approach based on bag-of-words (BoW) representation method that consists of a sequence of characters or words. There have been so many studies which use representations for the traditional text classification tasks. A multi-label text classification, a document can be belonging to more than one categorized to any of three classes: sports, politics, and movies. Hence classification of such documents belonging to more than one category can be achieved through multi-label text classification. In this project, we identify those documents which are strongly related to particular class from the stream of those documents. The document does not belong to any classes is kept in the residue. The problem of imbalances in the corpus is solved, where the documents should more or less equal in every class in the corpus. Represent each document in the BoW representation. In each class, the similar looking samples are clustered using SVM, KNN classifiers. For experimental purpose, two datasets are used viz., 20Newsgroup Mini dataset and Reuters. The classification results are validated with the help of precision, recall, F-measure and accuracy. To check the efficiency of the proposed model, the comparative analyses given against the other models. The results show that the proposed model outperforms the other models with respect to f-measure and accuracy.*

Keywords: 20Newsgroup Mini dataset and Reuters, Preprocessing, Text representation, SVM, KNN.

I. Preamble

A Text classification technique is necessary to find relevant information in many different tasks that deal with largest quantities of information in the text form. The World Wide Web is a huge source for storing and accessing linguistic documents. The amount of linguistic documents in the World Wide Web grows rapidly and this rapid growth has raised the great interest in helping people for finding easier ways to organize and classify these resources. Automatic text categorization or text classification methods enable the organization or categorization of a set of documents into different categories or classes. Many classification problems have been solved manually by the use of some rules commonly written by hand. But creating these rules is labor-intensive. Instead of using hand-written rules, the text categorization approaches uses machine learning methods to learn automatic classification rules based on human labeled documents. It is obvious that labeling is an easier task than writing rules. Hence, text categorization can be considered as an effective method for