

A NOVEL MODEL FOR TEXT DOCUMENT REPRESENTATION: APPLICATION ON OPINION MINING DATASETS

ASMAA MOUNTASSIR, HOUDA BENBRAHIM & ILHAM BERRADA

ALBIRONI Research Team, ENSIAS, Mohamed 5 University, Souissi, Rabat, Morocco

ABSTRACT

In this paper, we propose a novel model for Document Representation in an attempt to address the problem of huge dimensionality and vector sparseness that are commonly faced in Text Classification tasks. The proposed model consists of representing text documents in the space of training documents. To evaluate the effectiveness of our model, we focus on a problem of binary classification. We conduct our experiments on Arabic and English data sets of Opinion Mining. We use as classifiers Support Vector Machines (SVM) and k-Nearest Neighbors (kNN). We compare the performance of our model with that of the classical Vector Space Model (VSM) by the consideration of three evaluative criteria, namely dimensionality of the generated vectors, time taken by the classifiers, and classification results in terms of accuracy. Our experiments show that the effectiveness of our model depends on the used classifier. Results yielded by k-NN when applying our model are the same as those obtained when applying the classical VSM. For SVM, results yielded when applying our model are, in general, slightly lower than those obtained when using VSM. However, the gain in terms of time and dimensionality reduction is so promising since they are dramatically decreased by the application of our model.

KEYWORDS: Document Representation, Text Classification, Opinion Mining, Machine Learning, Natural Language Processing