Effective feature selection is a key component for building an efficient automatic document classifier. We regularly encounter in the Arabic literature—especially the scientific one—in-frequent non-Arabic words that are eliminated by practice during the pre-processing phase. Although infrequent, those words are highly pertinent to their documents and, thus, can contribute to build a more efficient classification model and enforce the subjectivity of the decision taken by the classifier. Therefore, we propose in this paper four different feature selection solutions that allow both preserving a maximum number of those words and getting satisfactory classification accuracy.
- István Pilászy. Text Categorization and Support Vector Machines. In the Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence
- Plantié M., Roche M., Dray G., EGC 2008 : Un système de vote pour la classification de textes d&apos;opinion, Laboratoire LGI2P, Laboratoire LIRMM.
- Raheel, S. L&apos;organisation des Connaissances et la Recherche d&apos;Information Textuelles par l&apos;Application des Méthodes Statistiques. 7ème colloque du chapitre français de l&apos;ISKO. Lyon, France.

Index Terms

Computer Science
Pattern Recognition
Keywords
Arabic Text mining  Machine Learning  Dimensionality Reduction  Automatic Classification