॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
INSTITUTE OF COMPUTER STUDIES
AND RESEARCH

CASE STUDY

Text Categorization with Classification Algorithms

BY,

Name: Shreyasi Patwardhan (18030142025)

Sagar Kelkar (18030142015)

Abstract

There are several researches and procedures for classifying Arabic language texts that are based mostly on different environments. This lack of dependence on a unified standard (such as a unified dataset) makes it hard to determine the most accurate technique for classification. In this paper, we study and analyze the classification algorithms based on a unified environment and a different dataset with the included challenges faced by these algorithms to demonstrate their effectiveness and accuracy with a large dataset. Keywords: Arabic Text Classification; Naive Bayes; Decision Tree; KNN

Introduction

A tremendous amount of pages and topics in the online world are becoming accessible to everyone. Users can easily write threads and upload files onto web pages giving people a great opportunity to share massive amounts of data. However, such advances gave rise to new challenges such as the ability to retrieve the required piece of information efficiently and effectively. Retrieving what the user wants or even the closest topics to the user's request, is at the core of the information retrieval discipline. What makes this problem complex and challenging is the large number of existing topics with overlapping terminology. In the world of computer and internet, there must be solutions to these problems; otherwise the process of searching and retrieving information on the internet is useless and may take a long time to reach the user request. Here comes the importance of the classification text in order to facilitate the retrieval of the required information. There are many areas for subjects such as medicine, sports, health, law, etc. Narrowing down the search space by focusing on the domain in which the user is interested is likely to improve the information retrieval process. The text classification is the automated technique used to classify the text in predefined category which is more related to the text. Part of the importance of text classification comes from its wide application. In addition to the traditional uses in information retrieval, other applications of text classification include spam filtering , sentiment analysis , determining author's characteristics such as identity , gender , dialect , native language , political orientation , etc. Most research has focused on classifying texts written in the English language. Other languages such as Arabic received less interest due the nature of these languages and the difficulty of their structures. The difficult nature in the Arabic language makes it more complex and difficult to deal with them because of the many rules and anomalous characteristics, but it has become necessary In this paper, we have studied many classification algorithms of Arabic language texts, there are many algorithms used for classification. we chose some of the text and classification algorithms and we have applied it is to the dataset written in Arabic language, each of these algorithms has the characteristics and standards, such as precision, Recall, F-measure and accuracy.

In this paper, we applied some algorithms in a different of Arabic dataset and make comparison between them to help to make the decision of what the algorithm that we will use and when can be used, depend on the results we have obtained from the comparison. Several algorithms, concerned the classification of texts, but differ in terms of accuracy in this paper that are interested in the process of classifying texts we have made a comparison between the algorithms for text classification in Arabic.

Methodology

Algorithms : KNN , Naïve Bayes, Decision Tree

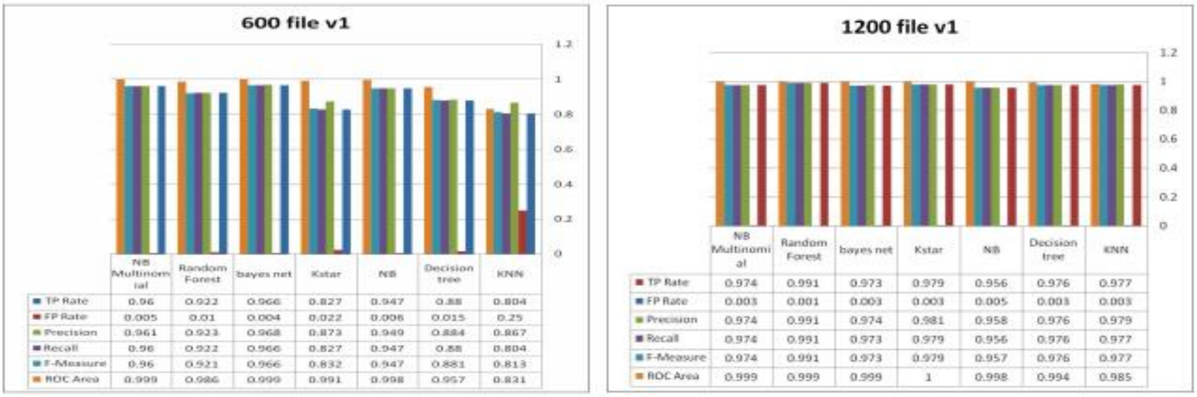
Most works have focused on classifying texts written in the English Language more than the Arabic language because of the Arabic nature and the difficulty of its structures. The difficult nature of the Arabic language makes it more complex and difficult to deal with because of the many rules and anomalous characteristics. However, it has become necessary to deal with this language because of its widespread usage online. To facilitate the search and retrieval in the Arabic language there are many algorithms working on the text classification that helps to retrieve data related to research in a short time and high accuracy. In this work we study many classification algorithms of Arabic language texts. In this paper, we choose some of the text and classification algorithms and apply them to the dataset written in Arabic language. Each of these algorithms have certain characteristics and standards, such as relative precision, recall, f-measure and accuracy. Our problem with this paper is to find when the algorithm is the best among the others depending on the results. There are several works and studies on text categorization of Arabic text and every work considers some points and leaves others depending on the type of study. In the authors consider classification of Arabic text that is very robust and reliable without morphological analysis. In [20] the authors conduct a comparative study using N-Gram and using two measures, Manhattan measure and Dice's measure. They compare them together and the result was that the N-Gram with Dice's measure is better than using Manhattan measure. In [35] considers both labeled and unlabeled documents using expectation-maximization (EM). They proposed an algorithm based on EM with the Naive Bayes (NB) classifier to learn from the documents, both labeled and non-labeled. The SVM classifier gave better accuracy when the number of selected word is small, but the performance of the KNN classifier outperforms the SVM classifier when the numbers of selected words increase. Both classifiers reach the 100% accuracy when the number of selected words equal 450, this is due to the lack of the sufficient number of training and testing documents, where 98% of the documents were used for training the classifiers and only 2% of the documents were used for testing.

Dataset

We use in this paper a dataset that is divided into five parts. Each part has nine categories: Art, Economy, Health, Law, Literature, Politics, Religion, Sport and Technology.

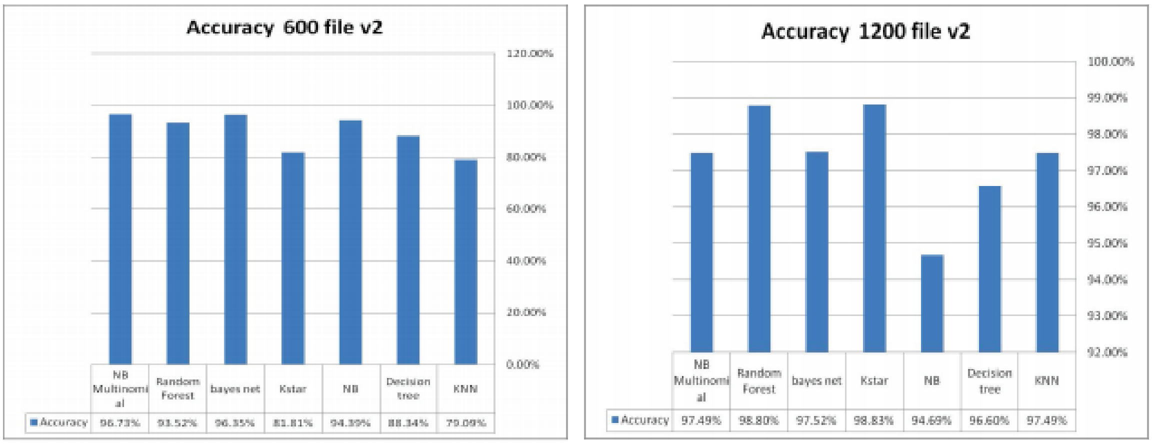
- Part 1: The original dataset without changes.
- Part 2: Dataset by removing stop words, punctuations and diacritics.
- Part 3: Dataset with applying the light 10 stemmer.
- Part 4: Dataset with applying Chen stemmer.
- Part 5: Dataset with applying Khuja algorithm for extracting the roots.

Classification algorithms were applied to each part, with applied features reduction. We experiment the algorithms using 600 and 1200 files for each category and then record the results for each experiment.



(a) 600 files

(b) 1200 files



(a) 600 files

(b) 1200 files

Conclusion:

In the text classification there are some algorithms concerned with Arabic text. We study these algorithms to determine which one is good. We applied the algorithm with five Parts of data and the results showed that the accuracy vary from one algorithm to another depending on the nature and size of data.

Random forest will give the highest accuracy according to the dataset with 1200 file.

References:

[1] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM TOIS, 26(3):12, 2008. [2] N. Abdulla, N. Mahyoub, M. Shehab, and M. Al-Ayyoub. Arabic sentiment analysis: Corpus-based and lexicon-based. In Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2013.

[2] N. Abdulla, N. Mahyoub, M. Shehab, and M. Al-Ayyoub. Arabic sentiment analysis: Corpus-based and lexicon-based. In Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2013.

[6] N. A. Ahmed, M. A. Shehab, M. Al-Ayyoub, and I. Hmeidi. Scalable multi-label arabic text classification. In Information and Communication Systems (ICICS), 2015 6th International Conference on, pages 212–217. IEEE, 2015.