ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 57-61 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp23.pdf

Arabic Text Classification Using Bayes Classifiers

Hadeel N. Alshaer
M.Sc. Computer Science. Amman Arab University
Amman/Jordan
HadeelAlshaer94@outlook.com
Bayan S. Alzwahrah
M.Sc. Computer Science. Amman Arab University
Amman/Jordan
Bayan.Alzwahrah@Yahoo.com
Mohammed A. Otair
Chairman of CIS Dept., Amman Arab University
Amman/Jordan
Otair@aau.edu.jo

***Abstract*— A huge number of documents are available at the webpages in every days. These documents should be marshaled automatically for its suitable uses in order to use them effectively. In order to decide which document belongs to a specific class, one of the existing techniques such as text classification should be used. Arabic text classification being one of the most important issue in information retrieval field because the privacy of Arabic language. In this paper, Bayes text classifier will be studied with its six variations that found in WEKA tool. The results of the classifiers were compared using three well-known measures: Precision, Recall and F-Measure.**

Keywords—*Arabic Text Classification; Naïve Bayes; Weka; Precision; Recall; and F-Measure.*

I. INTRODUCTION

With the technological development in the world of informatics, this has led to an increase in the volume of files day after day, and the increase in the use of search engines, making it difficult to retrieve files containing the titles required from the search so that the texts need to be classified.

Data classification is the method of organizing data into classes for its most effective and efficient use. A well-planned data classification system makes necessary data easy to find and retrieve. This can be of particular meaning for risk management, legal discovery, and compliance. Written measures and guidelines for data classification must define what classes and criteria the organization will use to classify data and specify the roles and responsibilities of staffs within the organization regarding data stewardship. Once a data-classification scheme has been created, security ethics that specify appropriate handling practices for each class and storage standards that define the data's lifecycle requirements must be addressed.

In order to facilitate the retrieval of files as required and within the content, and emerged many classification algorithms related to different types of data. This paper tries to study and compare the following Bayes Classifiers variations: *Bayes Net, Naïve Bayes, Naïve Bayes Multinomial, Naïve Bayes Multinomial Text, Naïve Bayes Multinomial Updateable and Naïve Bayes Updateable.* In order to experiment the classifiers, they were applied on the datasets that collected from various news articles, such as newspapers, magazines or websites. The datasets are categorized into five classes such: Economy, Local, Religions, Sports and Technological.

The Arabic language is the official language of the Arab world. It is the language of the Holy Quran and a language of fundamental value in the life of every nation. It carries many ideas and meanings. It is the basis of the Arab nation and protects it from extinction. Arabic is a rich sea of wide meanings, vocabulary, words and structures. It is one of the world's largest languages in terms of terms, meanings and structures.

The Arabic language belongs to the family of Semitic languages, which is part of the African and Asian languages group and is the official language of all Arab countries and is one of the six official languages of the United Nations. Arabic contains 28 characters:

(أ - ب – ت – ث – ج – ح – خ – د – ذ – ر – ز – س – ش – ص – ض – ط – ظ – ع – غ – ف – ق – ك – ل – م – ن – هـ - و – ي).

There are pre-test Dataset operations called Natural Language Processing (NLP), which is an Artificial Intelligence (or Machine Learning) field that is capable to understand human speech. It is one of the fields of computational computer science and linguistics concerned with interactions between computers and human languages. The challenges to understanding the natural language are to understand and analyze language.

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 57-61 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp23.pdf

II. PREVIOUS STUDIES

Cooper in [1], provides a natural and effective way to represent probability dependency among a range of variables. Therefore, many researchers explore the use of the belief that networks represent knowledge of artificial intelligence. The algorithms have been developed for effective probabilistic processes using special categories of Belief Networks.

Watson in [5], mentions that Naive Bayes is a great learning workbook with the assumption that the features are independent and although the independence is poor, Naive Bayes assumes that it competes with the most advanced Classifiers in the classification process. The main objective of Naive Bayes is to understand the characteristics of the data that affect its performance.

Alexander and Cheryl in [6], tried to experiment that the Naïve Bayes in the Presence of Imbalance Naive Bayes Multinomial is a classifier that is used in many applications extensively and successfully when applied to text classification. This classifier requires a form of Smoothing when estimating Parameters. Researchers have suggested many successful boot formats and also that there are pre-treatment techniques that have harmful effects when used.

Langley and Wayne in [4], said that one of the objectives of automatic learning is to discover the principals involved in algorithms and domain characteristics. To this end, researchers conducted systematic experiments with nature. Researchers also focused on theoretical analysis and often on the model of learning probability most empirical studies often rely only on informal analyzes of the learning task. Other approaches involving the formulation of the intermediate status models of qualitative algorithms and their testing through experimentation and work on Conjunctive Learning provide an excellent example of this technique, by assuming information about the concept of purpose.

In [2], George provides a simple approach with clear connotations for the representation, use and learning of probabilistic knowledge, this method is designed to be used under the supervision of ascetic tasks whose aim is to be accurate prediction any category depends on the training that you have received through training data.

III. THE PRE-PROCESS PHASE

The collected Arabic dataset needs to be pre-processed before implementing the classifiers in order to enhance the classifying results. This phase could be consider as a Natural Language Processing NLP that can be divided into four phases: *tokenization, normalization, stop word removal and stemming.*

Firstly, tokenization is a process of segmentation of a sequence of string into pieces of keywords, phrases, symbols, etc., called Tokens. Token can be single words or whole sentences in the Tokenization process and some characters, such as punctuation, are ignored. Tokens become inputs for other operations such as parsing and data mining. The challenges of tokenization depends on the type of language and each token is separated and the other depending on white spaces. Documenting and tokenization is done by separating each word from the other in the document, stripping the words and treating the letters of accents and pronouns as token. This is useful in the information retrieval process by clarifying the words for easy understanding by data mining tools such as WEKA.

The second phase is a normalization which is a process of converting a set of words into a more consistent and precise sequence. So that word processing will be easy as it converts words into a standard form through operations that make it capable of working with and processing data. Normalization improves text matching by taking into account synonyms of the word meaning, method of writing and abbreviations. This helps greatly in the process of retrieving data and information. In addition, normalization is an important process for the texts used in the process of retrieving information and improving the process for the best results.

The third phase treats a stop word which is a letter or word that does not mean meaning alone, but it has a role of connecting the sentences in order to complete the meaning. Stop word removal is one of the pre-processing steps for the language. It simply removes the stop word from the document. Stop word removal process removes the meaningless words such as pronouns. In other words, it removes the words that do not affect the meaning but improve the results of the retrieval and classification processes.

The last phase is the stemming which returns the words to their roots and the deletion of additional characters such as "وا", "ون", "و", "ال". Stemming is a pre-processing step of data processing for information retrieval and classification. Stemming usually refers to the removal of the ends of words in the hope of reaching the optimal solution correctly and accurately and often represents the removal of derivational affixes [3]. This process is used in the search engines to meet the user's needs for a particular word to be searched by its root, thus increasing the available options and then improving the results.

After implementing the above four phases, the datasets are converted to Attribute Relation File Format (ARFF). ARFF is a text file that describes a list of situations that share the same characteristics and common features of the original text. It has been developed by the Waikato University Learning Project for use with the WEKA Tool.

WEKA tool is machine learning workbench, a machine-learning Platform based on Waikato Environment for knowledge analysis. WEKA Tool has many features such as: open source, graphical interface, command line interface, Java API and documentation. The reason for the success of the WEKA is that can be used easily by the beginners who do not having an enough knowledge with any programming language.

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 57-61 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp23.pdf

Moreover, WEKA has several functionalities such as: data preprocessing, classification, clustering, attribute evaluators, search for features selection, algorithm for finding association rules and graphical user interface.

After loading the ARFF on WEKA, we activate the StringToWordVector Filter on ARFF where the filter extension is: Weka.Filters.Unsupervised.attribute.StringToWordVector, It is a filter that converts the String Attribute into a set of attributes to represent word occurrence.

The results of running of all the six Classifiers will be compared using a well-known measures such as: Precision, Recall and F-measure.

Precision is the positive predictive value. It is a part of the related instances between retrieved instances from the results of the process and is expressed by the following equation:

$$Precision = \frac{|\{\text{Relevant Document}\} \cap \{\text{Retrived Document}\}|}{|\{\text{Retrived Document}\}|}$$

The Recall is described as sensitivity measure, which is a part of the relevant instances of the over the total of number of the relevant instances and is expressed as:

$$Recall = \frac{|\{\text{Relevant Document}\} \cap \{\text{Retrived Document}\}|}{|\{\text{Relevant Document}\}|}$$

The third measure for comparison is F-measure. It is a measure of the accuracy of the tested classifier, which is the harmonic rate of the precision and recall and it could be expressed as follows:

$$F - measure = 2 * \frac{(\text{Precision} * Recall)}{(\text{Precision} + Recall)}$$

IV. Bayes Cllassifiers

A Bayes Net is a model that reflects states that are part of the world being modeled and describes how they relate to one another. The model may be linked to any Entity or States in this world and can be represented by Bayes Net and all existing and potential States that exist and can be modeled by the Bayes Net.

The possibilities come from the assumption that some States occur frequently when other states exist. This model is useful because it helps us to understand the world we want to model and helps us to predict the results of States in this world. This model is often easy to represent and model the world and the States From which.

Bayes Net refers only to a contract that is potentially linked to a kind of causality and can result in massive savings in calculations. There is no need to store all possible possessions from States and all that is required for storage and work with all possible groups of states among sets of nodes. This saves large areas of tables and accounts. Bayes Net helps with decision-making and helps to achieve the most accurate prediction process, simply balancing your own model with degrees of goodness and badness.

The main goal is to achieve the state maximize the pleasure and minimize the pain, which is called Bayes for the world Thomas Bayes. In the summary about Bayes Net it is a model and every model represents a state in this world and there is also a relationship between these States which is a mathematical tool for modeling the world which is flexible and adaptable to any degree of knowledge and is computationally efficient and can be applied anywhere.

Naïve Bayes is a classifier from the Bayes family and is based on the Bayes Theorem. Assuming independence between novices, the Naive Bayes model is easy to build and very useful with the large numbers, although it is simple to install and often outperforms its classifiers. It is an algorithm used for the classification process and uses the Spam Filters. The Bayes classifiers are widely used for automatic learning because they are easy to implement. Naïve Bayes Multinomial is a specialized version of Naive Bayes that is more precisely designed for text files. The task of classification is after the introduction of the training data. The workbook automatically classifies the categories we entered into the automatic learning process.

Naïve Bayes Multinomial Updateable is an extension version of Naïve Bayes Multinomial; whereas a Naïve Bayes Updateable is a modified version of Naïve Bayes Classifier.

V. Test Bayes Classifiers

In this section, six Bayes Classifiers will be tested using Weka and analyze their results via the standard measures that are: precision, recall and F-Measure. The classifiers were implemented on a specific dataset generated from the authors of the paper. The dataset contains 250 texts from several newspapers and articles, and classified into five classes as follow: Economy, Local, Religions, Sport, and Technological. Every class contains 50 texts, and the number of words in every texts are varied from 50 to 1000 words.

A. Bayes Net

The second experimented classifier is Bayes Net, table 1 shows the WEKA results for this classifier. This table and the remaining tables (from table 2 to table 6) contain four columns as follow: Class Name, Precision, Recall, and F-Measure. The last row is the average of the whole classes. The best resulted precision was with the Economy class, while the best recall was with the Religions class.

TABLE 1: BAYES NET RESULT

Classes	Precision	Recall	F-Measure
Economy	1.000	0.900	0.947
Local	0.933	0.840	0.884
Religions	0.893	1.000	0.943
Sport	0.907	0.980	0.942
Technological	0.980	0.980	0.980
Average	**0.943**	**0.940**	**0.939**

The range values of the results of the Bayes Net classifier were very high.

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 57-61 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp23.pdf

B. Naïve Bayes

The second experimented classifier is Naïve Bayes, table 2 shows the WEKA results for this classifier. The best resulted precision was with the Economy class, while the best recall was with the Religions and Sport classes.

TABLE 2: NAÏVE BAYES RESULT

Classes	Precision	Recall	F-Measure
Economy	1.000	0.940	0.969
Local	0.935	0.860	0.896
Religions	0.943	1.000	0.971
Sport	0.980	1.000	0.990
Technological	0.925	0.980	0.951
Average	**0.957**	**0.956**	**0.955**

The average values for Precision, Recall, and F-Measure using Naïve Bayes are higher than the resulted values Bayes Net classifier.

C. Naïve Bayes Multinominal

The third experimented classifier is Naïve Bayes Multinominal, table 3 shows the WEKA results for this classifier. The best resulted precision was with the Religions class, while the best recall was with the Sport classes.

TABLE 3: NAÏVE BAYES MULTINOMINAL

Classes	Precision	Recall	F-Measure
Economy	0.960	0.960	0.960
Local	0.885	0.920	0.902
Religions	1.000	0.880	0.936
Sport	0.962	1.000	0.980
Technological	0.942	0.980	0.961
Average	**0.950**	**0.948**	**0.948**

The results of Naïve Bayes Multinomial were high and so closed to the results from Naïve Bayes.

D. Naïve Bayes Multinominal Text

The fourth experimented classifier is Naïve Bayes Multinominal Text, table 4 shows the WEKA results for this classifier. All of the results were very low, except the recall value for the Economy class.

TABLE 4: NAÏVE BAYES MULTINOMINAL TEXT RESULT

Classes	Precision	Recall	F-Measure
Economy	0.200	1.000	0.333
Local	0.000	0.000	0.000
Religions	0.000	0.000	0.000
Sport	0.000	0.000	0.000
Technological	0.000	0.000	0.000
Average	**0.040**	**0.200**	**0.063**

E. Naïve Bayes Multinominal Updateable

The fifth experimented classifier is Naïve Bayes Multinominal Updateable, table 5 shows the WEKA results for this classifier. The best resulted precision was with the Religions class, while the best recall was with the Sport classes.

TABLE 5: NAÏVE BAYES MULTINOMINAL UPDATEABLE RESULT

Classes	Precision	Recall	F-Measure
Economy	0.960	0.960	0.960
Local	0.885	0.920	0.902
Religions	1.000	0.886	0.936
Sport	0.962	1.000	0.980
Technological	0.942	0.980	0.961
Average	**0.950**	**0.948**	**0.948**

All the results of Naïve Bayes Multinomial Updateable are matched exactly to the results of Naïve Bayes Multinominal classifier.

F. Naïve Bayes Updateable

The sixth experimented classifier is Naïve Bayes Updateable, table 6 shows the WEKA results for this classifier. The best resulted precision was with the Economy class, while the best recall was with the Religions and Sport classes.

TABLE 6: NAÏVE BAYES UPDATEABLE RESULT

Classes	Precision	Recall	F-Measure
Economy	1.000	0.940	0.969
Local	0.935	0.860	0.896
Religions	0.943	1.000	0.971
Sport	0.980	1.000	0.990
Technological	0.925	0.980	0.951
Average	**0.957**	**0.956**	**0.955**

All the results of Naïve Bayes Updateable are matched exactly to the results of Naïve Bayes classifier.

VI. COMPARISON THE SIX CLASSIFIERS

This section compares the results of the six Classifiers, in order to describe the best classifier among all the classifiers resulting from the classification process using Bayes Classifiers that were applied on an Arabic dataset. Table 7 summarizes all the results from tables 1 to 6. It takes the last row from all tables that represents the average values of: Precision, Recall, and F-Measure.

TABLE 7: CLASSIFIERS RESULT ANALYSIS

Classifier	Avg. Precision	Avg. Recall	Avg. F-Measure
Bayes Net	0.943	0.940	0.939
Naïve Bayes	0.957	0.956	0.955
Naïve Bayes Multinomial	0.950	0.948	0.948
Naïve Bayes Multinomial Text	0.040	0.200	0.063
Naïve Bayes Multinomial Updateable	0.950	0.948	0.948
Naïve Bayes Updateable	0.957	0.956	0.955

It is noticed that the best values were generated from Naïve Bayes and Naïve Bayes Updateable. In the other side, the worst results were from Naïve Bayes Multinomial Text.

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 57-61 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp23.pdf

VII. Future Work

In the near future, we want to improve on Bayes algorithms by combining the properties of all algorithms to access Hybrid Algorithm to solve the problems of other classification algorithms and to obtain the Best Classification Algorithm. We will try to study the algorithms to arrive at the best classification algorithm that works with the Arabic language to try to adopt them mainly in the specialized classification studies in Arabic.

References

[1] F. Cooper, "The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks", Journal Artificial Intelligence, Vol. 42, No.2, 1990.

[2] H. George, "Estimating Continuous Distributions in Bayesian Classifiers", UAI'95 Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995.

[3] M. Otair, “Comparative Analysis of Arabic Stemming Algorithms”. International Journal of Managing Information Technology (IJMIT), Vol. 5, No.2, 2013.

[4] P. Langley, and K. Wayne, "An Analysis of Bayesian Classifiers", In Proceedings Of The Tenth National Conference On Artificial Intelligence, 1995.

[5] T. Watson, "An empirical study of the naive Bayes classifier", IJCAI 2001 workshop on empirical methods in artificial intelligence.

[6] Y. Alexander, and E. Cheryl, "Smoothing Multinomial Naïve Bayes in The Presence of Imbalance", International Workshop on Machine Learning and Data Mining in Pattern Recognition, 2011.