First Libyan International Conference on Engineering Sciences & Applications (FLICESA_LA)
13 – 15 March 2023, Tripoli – Libya

Comparing the types of Naive Bayes in the Classification of the Arabic text

Fawzia Mansur
dept. Computer Science
Omar Al-Mukhtar University
Albayda,Libya
fawzia.abdalgani@omu.edu.ly

Abstract*— Arabic is one of the Semitic languages of antiquity. It is one of the six official languages of the UN. Arabic Classification plays an important and essential role in modern times Apps. There is a big difference between dealing with English text and Arabic text classification. preprocessing is also challenging for Arabic text. In order to determine to which, document it belongs A specific category, an existing technology such as text classification must be used. The classification of Arabic texts is one of them. One of the most important problems in the field of information retrieval for the confidentiality of the Arabic language. In this article, Bayes the text classifier will be studied in its six forms found in WEKA tool. The results of the classifiers were compared using three well-known metrics: precision, recall, and F-measure. Keywords— *Arabic Text Classification; Naïve Bayes; Weka; Precision; Recall; and F-Measure.

I. INTRODUCTION

Machine-readable information is available in large size and the increase in quantities complicates comprehension and use. Machine learning (ML) provides tools that help organize vast numbers of texts and detect them automatically [1]. Data classification is the method of organizing data in categories for its most effective and efficient use. Well planned data classification system makes it easy to find the necessary data and retrieve it. This can be of particular meaning for risk management, legal discovery, and compliance. Metrics and guidelines for data classification should be defined what categories and criteria will the organization use classify data and specify the roles and responsibilities of staffs within the organization regarding data stewardship. Once the data classification scheme has been established, the security ethics determine appropriate handling practices for each category and storage standards that define the data's lifecycle requirements must be addressed. In order to facilitate the recovery of files as needed and within the content, many classifications have emerged Algorithms related to different data types. 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. For a purpose experiment with classifiers applied to datasets extract from Al-Hayat News in Arabic. The datasets are categorized into seven categories such as: news, economy, science, automotive, technology, general and sports. 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-tested dataset processes called Natural Language processing (NLP), which is an Artificial Intelligence (or machine learning) that is capable of it understand human speech. It is one of the areas computational computing and linguistics involved with interactions between computers and human languages. The challenges to understanding the natural language are to understand and analyze language.

II. RELATED WORKS

In literature, several articles compare the Arabic text classification methods. Next, a comparison studies on the classification of Arabic text carried out over the last ten years: The authors, in [2], compare six classification models: Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Scission Tree (DT), Logistic Regression (LR). The authors used only one feature extraction technique, namely, the Bag of Words (B o W). For his experiences, the authors used the DSAC group [3]. Results are experienced Showed that logistic regression the best f-score. Watson in [4], says Naive Bayes is brilliant Learn the workbook assuming that the features Independent and, despite the lack of independence, naïve Bayes assumes it is competing with the more advanced
Workbooks in the classification process. The main objective of Naive Bayes is an understanding of properties of data that affect its performance. Alexander and Cheryl in [5], tried an experiment that Knife Bayes in the Presence of Naive Bayes Imbalance A polynomial is a classifier used in many applications widespread and successful when applied to text classification. This workbook requires some form of smoothing when estimating the parameters.

III. THE PRE-PROCESS PHASE

The Arabic dataset collected needs to be preprocessed

before running classifiers for optimization ranking results. This step can be considered normal NLP language processing which can be divided into four Steps: tokenization, normalization, stop word removal and stemming.

First, tokenization is a process of segmentation of a sequence of string into pieces of keywords, phrases, symbols, etc., called Tokens. The Token can be single or whole words phrases in the Tokenization process and some characters, Like punctuation, it is ignored. Tokens become entries for other operations such as analysis and data mining. The tokenization challenges depend on the type of language each token is separated and the other depending on white spaces. Documentation and tokenization are done per class each other's word in the document, abstract words and treat accented letters and pronouns as code. This useful in the information retrieval process by clarifying words for easy understanding by data mining tools such as WEKA.

The second step is normalization, which is a process convert a group of words into a more coherent and precise group 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 step deals with the 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 steps in preprocessing, it simply removes the stop word from the document. Stops the process of deleting words deletes meaningless words like 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 step is the derivation process which returns the words to their roots and remove extra characters like "ال", "و", "ون", "وا". Stemming is a data pre-processing step Information retrieval processing and classification. Stemming usually refers to the removal of word endings by correctly thinking about the optimal solution and accurately and often represents the removal of derivational affixes [6]. 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 steps, the datasets are converted to Attribute Relationship File Format (ARFF). The ARFF is a text file describing a list of posts that share the same thing. Characteristics and common characteristics of the original text. It has been developed by the Waikato University Learning Project for use with the WEKA Tool.

WEKA Tool is a machine learning classifier, which is a machine learning platform based on the Waikato environment for Knowledge analysis. WEKA Tool has many features such as: Open source, GUI, command line interface, java APIs and documentation. Reason for success WEKA is that it can be easily used by beginners who don't have sufficient knowledge of any programming language. In addition, WEKA has many functions such as: data preprocessing, classification, grouping, isotropy, Feature selection search, correlation search algorithm Rules and GUI. 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 will run every six seeders comparison using known measures such as: accuracy, reminder and measurement F.

Precision is the positive predictive value. It's part of Related cases among the cases extracted from the results process and is expressed by the following equation:

$$Precision = \frac{|\{Relevant\ Document\} \cap \{Retrived\ Document\}|}{|\{Retrived\ Document\}|} \quad (1)$$

Recall is described as a measure of sensitivity, which is part of the relevant cases of more than a total number. The relevant cases are expressed as follows:

$$Recall = \frac{|\{Relevant\ Document\} \cap \{Retrived\ Document\}|}{|\{Relevant\ Document\}|} \quad (2)$$

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

$$F\text{-}measure = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (3)$$

IV. BAYES CLASSIFIERS

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. This saves a large percentage and accounts. Bays net helps decision making helps achieve more accuracy prediction case, just compare your model with degrees of goodness and badness. The main goal is to reach the state of maximizing pleasure and minimizing pain, what Thomas Bayes called Bayes. In 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 Multinomials 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 analyzed their results with standard measurements. They are: Precision, Recall, Measure and. classifier where it was performed on a specific dataset generated by the authors out of paper. The dataset contains 340 texts from
Al-Hayat News Arabic newspaper, divided into seven categories Follow: news, economy, science, automotive, technology, general and sports. Each chapter contains 50 texts.

A. Bayes Net

The first classifier tried is Bayes Net, Table 1 displays 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 result recall was with News class.

TABLE 1: BAYES NET RESULT

Class	Precision	Recall	F-Measure
Cars	0.357	0.543	0.431
Computer	0.167	0.073	0.102
Economy	0.250	0.120	0.162
General	0.417	0.510	0.459
News	0.308	0.560	0.397
Science	0.500	0.480	0.490
Sports	0.391	0.188	0.254
Average	0.345	0.359	0.333

Range values for Bayes Net classifier results were close

B. Naïve Bayes

The second classifier tried is Naïve Bayes, Table 2 displays WEKA results for this classifier. best result accuracy recall was with News and Science class.

TABLE 2: Naïve BAYES RESULT

Class	Precision	Recall	F-Measure
Cars	0.321	0.391	0. 353
Computer	0.308	0.098	0.148
Economy	0.167	0.100	0.125
General	0.615	0.163	0.250
News	0.333	0.780	0.467
Science	0.574	0.780	0.661
Sports	0.351	0.271	0.306
Average	0.384	0.377	0.336

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

C. Naïve Bayes Multinomial

The third experimented classifier is Naïve Bayes Multinomial, table 3 shows the WEKA results for this Classifier, best result Accuracy recall was with Sports class.

TABLE 3: Naïve Bayes Multinomial RESULT

Class	Precision	Recall	F-Measure
Cars	0.571	0.087	0.151
Computer	0.333	0.024	0.045
Economy	0.231	0.500	0.316
General	0.485	0.327	0.390
News	0.250	0.060	0.097
Science	0.385	0.300	0.337
Sports	0.250	0.688	0.367
Average	0.356	0.290	0.249

The results of Naïve Bayes Multinomial were closed to the results from Naïve Bayes

D. Naïve Bayes Multinomial Text

The fourth experimented to try is Naïve Bayes multinomial 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 Multinomial Text RESULT

Class	Precision	Recall	F-Measure
Cars	?	0.000	?
Computer	?	0.000	?
Economy	0.150	1.000	0.260
General	?	0.000	?
News	?	0.000	?
Science	?	0.000	?
Sports	?	0.000	?
Average	?	0.150	?

E. Naïve Bayes Multinomial Updateable

The fifth experimented classifier is Naïve Bayes Multinomial Updateable, table 5 shows the WEKA results. best result Accuracy recall was with Sports class.

TABLE 5: Naïve Bayes Multinomial Updateable RESULT

Class	Precision	Recall	F-Measure
Cars	0.571	0.087	0.151
Computer	0.333	0.024	0.045
Economy	0.231	0.500	0.316
General	0.485	0.327	0.390
News	0.250	0.060	0.097
Science	0.385	0.300	0.337
Sports	0.250	0.688	0.367
Average	0.356	0.290	0.249

F. Naïve Bayes Updateable

The sixth experimented classifier is Naïve Bayes Updateable, table 6 shows the WEKA results for this classifier. best result Accuracy recall was with News and Science class.

TABLE 6: Naïve Bayes Updateable RESULT

Class	Precision	Recall	F-Measure
Cars	0.321	0.391	0.353
Computer	0.308	0.098	0.148
Economy	0.167	0.100	0.125
General	0.615	0.163	0.258
News	0.333	0.780	0.467
Science	0.574	0.780	0.661
Sports	0.351	0.271	0.306
Average	0.384	0.377	0.336

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	0.345	0.359	0.333
Naïve Bayes	0.384	0.377	0.336
Naïve Bayes Multinomial	0.356	0.290	0.249
Naïve Bayes Multinomial Text	?	0.150	?
Naïve Bayes Multinomial Updateable	0.356	0.290	0.249
Naïve Bayes Updateable	0.384	0.377	0.336

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.

VI. CONCLUSION

In this article, Bayes the text classifiers studied in its six variations that found in WEKA tool. The results compared using classifiers Three well-known scales: precision, recall, and the F-measure. The best values were generated from Naïve Bayes and Naïve Bayes Updateable. And the worst results were from Naïve Bayes Multinomial Text.

REFERENCES

[1] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, no. 6, p. e01802, 2019.
[2] M. A. H. Madhfar and M. A. H. Al-Hagery, "Arabic text classification: A comparative approach using a big dataset," in 2019 International Conference on Computer and Information Sciences (ICCIS), 2019, pp. 1-5.
[3] M. Biniz, "DataSet for Arabic Classification," Mendeley Data, V2, doi: 10.17632/v524p5dhpj.2, 2018.
[4] T. Watson, "An empirical study of the naive Bayes classifier", IJCAI 2001 workshop on empirical methods in artificial intelligence.
[5] 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.
[6] M. Otair, "Comparative Analysis of Arabic Stemming Algorithms". International Journal of Managing Information Technology (IJMIT), Vol. 5, No.2, 2013.
[7] M. Bilal, H. Israr, M. Shahid and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques," Journal of King Saud University - Computer and Information Sciences, 2015.
[8] O. Einea, Elnagar, A., & Al Debsi, R., "SANAD: Single-Label Arabic News Articles Dataset for Automatic Text Categorization.," Mendeley Data, V2, doi: 10.17632/57zpx667y9.2., 2019
[9] T. Zerrouki, "Tashaphyne, Arabic light stemmer.", 2019.
[10] M. Ahmed and R. Elhassan, "Arabic text classification review," International Journal of Computer Science and Software Engineering, vol. 4, pp. 1--5, 2015.