International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

The Impact of Pre-Processing Techniques on Arabic Text Categorization Using Logistic Regression

Muner Mosbah[1], Abdelatti Blg[2], Salih Qarash[3]
[1, 2]Computer Science Department, Faculty of Science, University of Zawia, Alajilat, Zawia, Libya
[3]General Electrical Company, Algmial, Libya
Email address: m.athaba@zu.edu.ly, a.blg@zu.edu.ly

***Abstract*—*In this research paper, the Text classification is an important topic. There is an immense volume of electronic documents available online. Text classification aims to classify documents into a set of predefined categories Arabic text classification is one application of Natural Language Processing (NLP It has been employed for the analysis and categorization of Arabic text. Text analysis has become integral in our lives due to the escalating volume of textual data, turning text classification into a big data challenge. Arabic text classification systems play a crucial role in preserving essential information across various domains, including education, health, and public services. In this research, Logistic Regression (LR) is investigated in Arabic text dataset. To the best of our knowledge, LR was rarely used for Arabic text data before. The dataset has seven main categories: news, economy, cars, computers, science, sports, and generals. Data collected was cleaned from Latin characters, numbers, and punctuation and stop words to investigate the effectiveness of the dataset. The dataset undergoes Arabic text preprocessing to address the unique characteristics of Arabic text. The experimental findings from this research demonstrate that the classification accuracy results are evidently influenced by the pre-processing steps.*

***Keywords*—*Logistic Regression, Arabic Text Categorization, pre-processing, TF-IDF, Arabic Document Classification.*

I. INTRODUCTION

Text categorization (TC) involves automatically assigning newly discovered texts to predefined groups, such as law, sports, economics, religion, and computer science. There are two sorting options: supervised and unsupervised. This study opts for supervised classification due to its noted superior accuracy. This method involves training the classification system by providing it with a set of already labeled text documents. Navigating through a vast and constantly expanding pool of internet textual material to find relevant information on a specific topic poses a significant challenge. Organizing data into predetermined categories can provide a valuable solution to address this issue. Text categorization algorithms play a crucial role in various natural language processing applications, including text description, query response, spam detection, and visualization [1].

Numerous algorithms have been developed to tackle the problem of text classification (TC).While a significant amount of research in this field has been dedicated to English text, comparatively less attention has been given to Arabic text. This is primarily due to the differences in morphological structure between English and Arabic, making pre-processing Arabic text more complex. The objective of this study is to assess the effectiveness of the Logistic Regression (LR) classifier for Arabic text classification. The performance of the classifier is evaluated based on precision, recall, and F1-measure.

II. RELATED STUDIES

In this section, the focus will be solely on researchers who have conducted studies using an Arabic dataset. While it's true that the volume of research conducted with Arabic datasets pales in comparison to those using English datasets, there are a handful of researchers who have successfully carried out experiments with Arabic datasets.

(Adel Hamdan et al. in [2] conducted several experiments using various algorithms such as Naïve Bayesian, K-Nearest Neighbours, Neural Network, Rocchio classifier, C4.5, and Support Vector Machine in their research. These experiments were carried out using an Arabic dataset. However, it's worth noting that the dataset used in these studies was smaller compared to the one used in this research.

Elnagar et al. in [4] introduced two new extensive corpora for Arabic text categorization, SANAD and NADiA, which were collected from news portals. The results demonstrated robust performance of all models on the SANAD corpus, with the lowest accuracy being 91.18%. For the NADiA corpus, the attention-GRU model achieved the highest overall accuracy of 88.68%.

Boukil et al. in [7] proposed a straightforward and accurate method for categorizing an Arabic dataset. They utilized an Arabic stemming algorithm to isolate, select, and reduce features. For feature weighting, they employed Term Frequency Inverse Document Frequency (TFIDF). They analyzed their dataset using the CNN model and other conventional machine learning techniques as a benchmark. They found that the CNN model performed well in the Arabic text classification challenge. Traditional methods, such as SVM, did not perform as well as the CNN model, particularly when dealing with large datasets. ABICABIC TEXT CATEGORIZATION (TC). Constructing a text categorization (TC) system typically initiates with a preprocessing phase to ready texts for automatic categorization, proceeds to the classifier training phase, and culminates in testing the classifier and assessing its performance using formal evaluation criteria. Arabic texts undergo certain pre-processing steps common to texts in other languages, such as removing stop words, stemming, and feature weighting and selection. However, owing to the unique characteristics of the Arabic language, additional specialized PRE-PROCESSING steps

Muner Mosbah and Salih Qarash, "The Impact of Pre-Processing Techniques on Arabic Text Categorization Using Logistic Regression," *International Research Journal of Advanced Engineering and Science*, Volume 8, Issue 4, pp. 128-131, 2023.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

are required. The specifics of the data PRE-PROCESSING stage employed in our research are elucidated in the subsequent subsection.

a) The Arabic language features and challenges

Arabic, an ancient Semitic language, boasts a deeply ingrained and well-established morphology, making it one of the most highly inflected languages. Consequently, a single word can convey an entire sentence through sequential concatenation [6]. Over 250 million people worldwide speak the Arabic language. Furthermore, as the language of the Holy Quran, more than a billion Muslims understand it. The Arabic alphabet comprises 28 characters: (ي و ه ن م ل ك ق ف غ ع ظ ط ض ص ش س ز ر ذ د خ ح ج ث ت ب أ).

III. PROPOSED CLASSIFICATION SYSTEM

The aim of text classification is to create a model that used to classify different text documents to its predefined classes. Figure 1 represents the classification model phases [9]. The development of a text classification system involves several stages. It starts with text preprocessing, then moves on to feature extraction, and finally employs machine learning to allocate the text to one of the predefined categories.

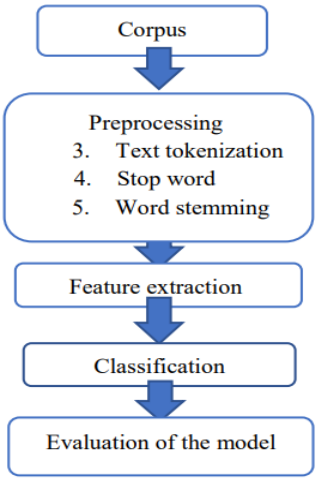

Fig. 1. Classification General phases

- Selected Classifier (LOGISTIC REGRESSION (LR) ALGORITHM)

Logistic Regression (LR) is a machine learning method that is often used for classification problems. It's a relatively straightforward technique that is widely employed. Logistic regression is a statistical approach designed to predict binary classes, where the dependent variable follows a Bernoulli distribution [11]. The logistic function, also known as a sigmoid function, is an 'S' shaped curve that maps any real-valued number to a value between 0 and 1. If the curve approaches positive infinity, a prediction of 1 is expected, and if it approaches negative infinity, a prediction of 0 is expected [10]. If our ultimate objective is classification, then given a test example x, Logistic Regression can directly estimate the conditional probability of assigning a class label y to the example by: T 1 P(y|x)= 1 + exp(-y x) α.

IV. THE DATASET

A significant challenge in text classification for both English and Arabic languages is the lack of a universally available dataset that can serve as a benchmark. There isn't a general Arabic dataset that different authors can use as a benchmark. Most researchers in Arabic text classification create their own datasets. Our dataset is composed of 350 Arabic news documents. We gathered data from various Arabic websites worldwide using the HTTRACK web crawler to build our corpus. Following this, we performed a cleaning and normalization process to retain only the Arabic text. Since each website has different categories, we opted to construct datasets based on categories rather than websites. Consequently, we selected the most common categories, namely News, Economic, Science, Cars, Technology, General, and Sport.

V. EXPERIMENTS AND ANALYSIS

In this research (LR) was used in text classification. Documents in these experiments belong to 7 categories. Documents are collected from several resources. The most familiar evaluation measures used in text classification are precision, recall and F1- measure these three measures are known to be reliable evaluation measures of the classifier effectiveness and have been used widely in evaluating Arabic TC systems.

A. Data Pre-processing

Data pre-processing is the first stage in building TC systems for all languages. Data-preprocessing aims to reduce the number of features used in building classifiers, thus reducing requirements of memory. We split our constructed dataset into seven classes as showen in the fig (2): News, Economic, Science, Cars, Technology, General and Sport. Each class has 50 documents (40 for training and 10 for testing)

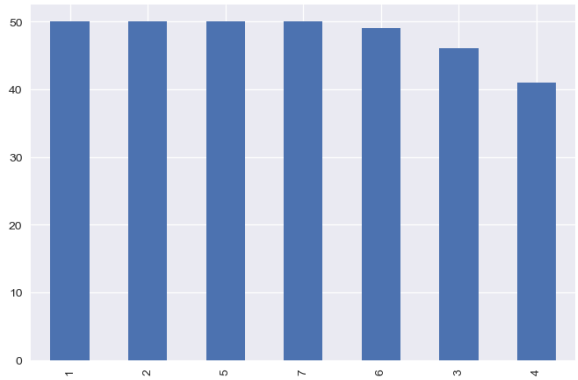

Fig. 2. The number of articles in each class

Furthermore, text pre-processing is used to clean the dataset by removing all the non-Arabic content. This approach is highly recommended when dealing with text collected from the web.

Stop Word Removal: Words that occur very frequently in the text and do not carry much meaning are removed from the text.

Muner Mosbah and Salih Qarash, "The Impact of Pre-Processing Techniques on Arabic Text Categorization Using Logistic Regression," *International Research Journal of Advanced Engineering and Science*, Volume 8, Issue 4, pp. 128-131, 2023.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

The next step is to clean all the scraped articles by removing stop words as showed below in fig (3).

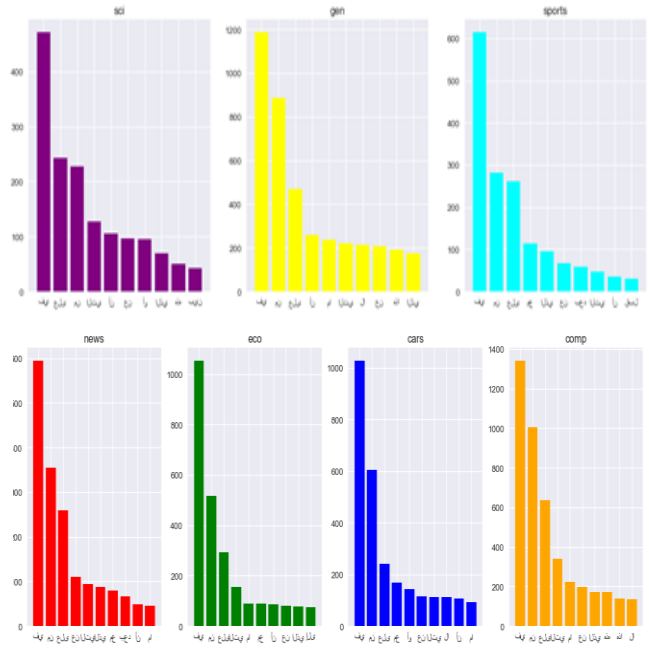

Fig. 3. The number of stop words in each class.

Punctuation Mark Removal: All the punctuations from the text are removed as they may not add much meaning to the text being processed, fig (4) showed the number punctuation in each class

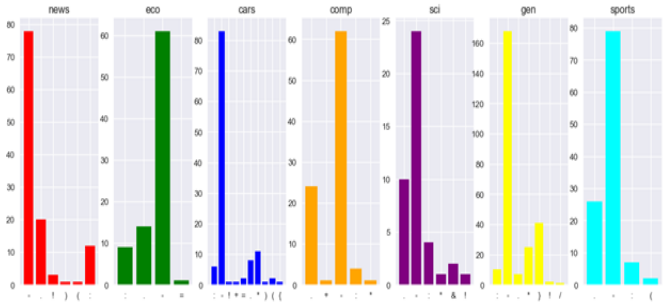

Fig. 4. The number of punctuation in each class

B. *Text Features*

In text processing, the words in articles represent categorical features. However, most machine learning algorithms cannot process text directly. To address this issue, we convert the text into numerical vectors, with each sentence represented by a single vector. This process of transforming text into vectors is known as vectorization. The most commonly used techniques for this task are Count Vectorizer and TF-IDF Vectorizer. Classifiers that use the TF-IDF Vectorizer tend to yield higher accuracies, which is why we chose to use it in our work. The term frequency-inverse document frequency (TF-IDF) reduces the impact of tokens that appear very frequently. The TF-IDF Vectorizer consists of two components:

- Term Frequency (TF): This measures how frequently a word appears in an article.
- as shown in fig (3). Since every article is different in length, it is possible that a term would appears much more times in long articles that shorter ones • Inverse Document Frequency (IDF): measures how important a word is by weighing

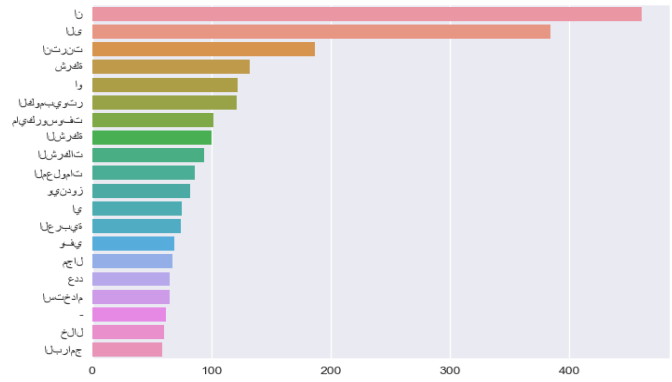

Fig. 5. The number of common words in all classes.

VI. Results And Discussion

Results of classifying the dataset using Logistic Regression with different pre-processing steps were conducted. The next table summarized different matrices results with pre-processing steps:

- Removal of url,html, special characters and punctuations.
- Removal of stop words.
- Stemming.

TABLE I. Metrics report with all pre-processing steps

Class	Precision	recall	F1 score	Support
News	0.83	0.91	0.87	11
Eco	0,93	1.00	0.97	14
Cars	1.00	0.83	0.91	6
Comp	1.00	1.00	1.00	9
Sci	0.92	1.00	0.96	12
Gen	0.86	0.67	0.75	9
Sports	1.00	1.00	1.00	7
Accuracy			0.93	68
Macro avg	0.94	0.92	0.92	68
Weighed avg	0.93	0.93	0.92	68

The next table summarized different matrices results with follow pre-processing steps:

- Removal of url,html, special characters and punctuations.
- Removal of stop words.
- Stemming not used.

TABLE 2. Metrics report with non-stemming used

Class	precision	recall	F1 score	support
News	0.80	0.73	0.76	11
Eco	1.00	1.00	1.00	14
Cars	1.00	1.00	1.00	6
Comp	1.00	1.00	1.00	9
Sci	1.00	0.92	0.96	12
Gen	0.64	0.78	0.70	9
Sports	1.00	1.00	1.00	7
Accuracy			0.91	68
Macro avg	0.92	0.92	0.92	68
Weighed avg	0.92	0.91	0.91	68

- The next table summarized different matrices results without any pre-processing steps:

Muner Mosbah and Salih Qarash, "The Impact of Pre-Processing Techniques on Arabic Text Categorization Using Logistic Regression," *International Research Journal of Advanced Engineering and Science*, Volume 8, Issue 4, pp. 128-131, 2023.

- × Removal of url,html, special characters and punctuations.
- × Removal of stop words.
- × Stemming not used.

TABLE 3. Metrics report without preprocessing steps

Class	Precision	recall	F1 score	support
News	0.92	1.00	0.96	11
Eco	0,90	0.92	0.91	14
Cars	1.00	1.00	1.00	6
Comp	1.00	1.00	1.00	9
Sci	1.00	0.92	0.96	12
Gen	0.78	0.75	0.77	9
Sports	0.88	1.00	0.93	7
Accuracy			0.90	68
Macro avg	0.92	0.93	0.91	68
Weighed avg	0.92	0.92	0.91	68

Results of classifying the dataset using Logistic Regression with different pre-processing steps were conducted table 1 displays the matrices results with accuracy 93% obtained by the classifiers with the using all preprocessing methods.". While the TABLE2 summarized matrices results with accuracy 91% with the use of only two pre-processing steps: Removal of url,html, special characters ,punctuations, Removal of stop words and Stemming not used. Finally, TABLE3 summarized matrices results with accuracy 90% achieved by the classifiers without using any preprocessing methods. We discovered that the used pre-processing techniques influence the accuracy the accuracy of the model.

VII. Conclusion

In this study, we explore the application of Logistic Regression (LR) for Arabic Text Categorization (TC). We conducted experiments on the widely-used Arabic Text Categorization Dataset and evaluated the classifier by calculating the precision, recall, and F1-measure. These measures are recognized as reliable indicators of classifier effectiveness and have been extensively used in evaluating Arabic TC systems. The LR classifier demonstrated highly accurate performance, these results suggest that LR is a promising and competitive algorithm for Arabic TC. It's clear from the results that the pre-processing steps significantly influence the classification accuracy, as shown in the tables.

Moreover, this classifier can be used for larger datasets successfully as long as good Feature Selection and Reduction criteria are applied on the dataset.

References

[1] K. Sundus, F. Al-Haj, and B. Hammo, "A Deep learning approach for Arabic text classification," 2019 2nd Int. Conf. New Trends Comput. Sci. ICTCS 2019 - Proc., pp. 1–7, 2019, doi: 10.1109/ICTCS.2019.8923083

[2] Adel, H., Tariq, A., & Omar, A. (2016). Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network. GSTF Journal of Computing, 5(1), 108-115.

[3] Mohammad, A.H., 2019. Arabic text classification: A review. Modern Applied Science, 13(5), pp.1-88.

[4] A. Elnagar, R. Al-Debsi and O. Einea, "Arabic text classification using deep learning models," Information Processing & Management, vol. 57, p. 102121, 2020.

[5] Sabri, T., El Beggar, O. and Kissi, M., 2022. Comparative study of Arabic text classification using feature vectorization methods. Procedia Computer Science, 198, pp.269-275.

[6] A. El Kah and I. Zeroual, "0e effects of pre-processing techniques on Arabic text classification," International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 1, pp. 41–48, 2021.

[7] S. Boukil, M. Biniz, F. El Adnani, L. Cherrat, and A. E. El Moutaouakkil, "Arabic text classification using deep learning technics," Int. J. Grid Distrib. Comput., vol. 11, no. 9, pp. 103– 114, 2018, doi: 10.14257/ijgdc.2018.11.9.09

[8] Alzanin, S.M., Azmi, A.M. and Aboalsamh, H.A., 2022. Short text classification for Arabic social media tweets. Journal of King Saud University-Computer and Information Sciences, 34(9), pp.6595-6604.

[9] Abdulghani, F.A. and Abdullah, N.A., 2022. A survey on Arabic text classification using deep and machine learning algorithms. Iraqi Journal of Science, pp.409-419.

[10] Arista, A., 2022. Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19. Sinkron: jurnal dan penelitian teknik informatika, 7(1), pp.59-65.

[11] Majumder, A. B., Gupta, S., Singh, D., & Majumder, S. (2021). An intelligent system for prediction of COVID-19 case using machine learning framework-logistic regression. Journal of Physics: Conference Series, 1797(1). https://doi.org/10.1088/1742-6596/1797/1/012011