University of Zawia Journal of Engineering Sciences and Technology (UZJEST)
Volume 1, Issue 1, (2023), pp 22-28, ISSN 3006-2764

Impact Feature Vectorization Methods on Arabic Large Data Using Logistic Regression Classification

Ali Shafah[1], Ahmed Suleiman[2], Samira Alshafah [2]

[1] Data Analysis Department, Faculty of Economics, University of Zawia, Zawia, Libya

[2] Computer Department, Faculty of Education, University of Zawia, Zawia, Libya

*Corresponding author email: ali.shafah@zu.edu.ly

Received: 09.11.2023 | Accepted: 19.11.2023 | Available online: 15.12.2023 | DOI:10.26629/uzjest.2023.03

ABSTRACT

The process of assigning text documents to a predetermined set of categories is known as text categorization. The objective of this study is to present experimental assessments of various feature vectorization methods for the purpose of categorizing a large Arabic corpus using a logistic regression classifier. N-Gram, Bag of Words, and Term Frequency–Inverse Document Frequency are these methods. A corpus of around 111,000 Arabic documents was utilized, which was split up into five categories: news, sports, culture, economics, and varied. Each method's experimental findings were assessed using three different performance indicators. According to the experimental findings, the Logistic Regression model using Term Frequency–Inverse Document Frequency and N-gram (1,2) had the best accuracy, scoring 96%, while Bag of Words came in second with 95%.

Keywords: Arabic Text Classification, Large data, Big data, Feature Vectorization, TF-IDF, BoW, N-gram.

تأثير طرق تحويل السمات على البيانات العربية الكبيرة باستخدام تصنيف الانحدار اللوجستي

علي الشفح[1]، أحمد سليمان[2]، سميرة الشفح[2]

[1]قسم تحليل البيانات ، كلية الاقتصاد الزاوية ، جامعة الزاوية ، الزاوية، ليبيا

[2]قسم الحاسوب، كلية التربية الزاوية، جامعة الزاوية، الزاوية، ليبيا

ملخص البحث

تُعرف عملية تعيين المستندات النصية لمجموعة محددة مسبقًا من الفئات باسم تصنيف النص. الهدف من هذه الدراسة هو تقديم تقييمات تجريبية لطرق تحويل سمات مختلفة بغرض تصنيف مجموعة كبيرة من النصوص العربية باستخدام مصنف الانحدار اللوجستي. **N-Gram**، و **Bag of Words** ، و **Term Frequency–Inverse Document Frequency** هي هذه الطرق. تم استخدام حوالي 111.000 وثيقة عربية، تم تقسيمها إلى خمس فئات: الأخبار والرياضة والثقافة والاقتصاد والمتنوعة. تم تقييم النتائج التجريبية لكل طريقة باستخدام ثلاثة مؤشرات أداء مختلفة. ووفقاً للنتائج التجريبية، فإن نموذج الانحدار اللوجستي باستخدام **Term Frequency–Inverse Document Frequency**

و**N−gram** (1,2) حصل على أفضل دقة، حيث سجل 96%، في حين جاء **Bag of Words** في المركز الثاني بنسبة 95%.

الكلمات الدالة: تصنيف النصوص العربية، البيانات الضخمة، البيانات الضخمة، متجهات الميزات، TF-IDF، BoW، N-gram

1. Introduction

Text classification is a crucial task in Natural Language Processing (NLP) that is utilized to tag and classify text according to its content. Large volumes of textual content can be categorized in order to organize the platform, improve and expedite automated navigation overall, and make information searches much simpler and more useful [1]. Many repositories are accessible online as a result of the increased usage of the Internet, which raises the need for automatic classification algorithms. Even though text-based and unstructured data makes up over 80% of data [2], it is nevertheless regarded as a very important and comprehensive source of knowledge.

Today, since there are so many text documents, expert manual classification has not proven as efficient. As a result, it was discovered that automated classifiers using machine learning techniques were superior and a great substitute. The use of text categorization has been investigated in numerous contexts and situations, including sentiment analysis [3-7], spam filtering, language identification, dialect identification [8-10], and many others. In the business world, machine learning is particularly useful for organizing data. It improves decision-making and automates procedures for quicker outcomes. For instance, marketers can investigate, gather, and examine the keywords utilized by rivals.

Considering the number of Internet users, Arabic language speakers make up the language group on the web with the greatest growth rate, with a population of just over 440 million. Approximately two-thirds of young Arabs, according to a survey, acquire their news online [11].

Understanding the syntactic structure of the words is necessary for creating a classification system for the Arabic language so that we may modify and represent the words more precisely. Comparatively less research has been done on classifying Arabic texts as opposed to English texts. Even less research has been done on Arabic short-text classification. Another reason is unrelated to the characteristics of the language [12].

Text processing in Arabic is particularly challenging due to the unique features of the language. These characteristics include 1) the language's inflectional and derivational structure, 2) the absence of capital letters and short vowels, 3) the existence of various dialects of the language in use, including modern, classical, and colloquial, and 4) the ability to construct sentences with implicit subjects [13].

In this paper, a Large Arabic dataset, collected by [14] was used, The dataset consists of 111,728 documents and 319,254,124 words. as well as studying the impact of feature vectorization methods on the performance of the logistic regression classifier.

This paper is structured as follows: Section 2 presents related works about logistic regression and feature vectorization methods. The experiment is presented in Section 3. Results and the discussion of these experiments are presented in Section 4. Finally, conclusions and future work take place in Section 5.

2. Related Work

The majority of text classification studies are created with English and other languages like German, Italian, and Spanish. Nevertheless, there aren't many studies done on Arabic language classification. A number of more recent studies have been proposed among those works which focus in logistic regression.

Tarik Sabri et al. [15] used five classification models using two Arabic datasets cnn_arabic and osac_uft8 to do an empirical study. These algorithms are Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN) and Logistic Regression (LR). Three feature vectorization methods were applied to convert text into numeric vectors word count, Terms Frequency-Inverse Document Frequency (TF-IDF) and word embedding using word2vec. For the applied feature vectorization techniques, the experiment shows that the classifiers SVM and LR score the highest performance followed by RF, KNN and DT. Besides, the experiment shows that feature vectorization methods and dataset size have a high impact on the performance of the algorithms RF, KNN and DT, while SVM and LR maintain stable outcomes.

Mayy M. Al-Tahrawi [16] investigated Logistic Regression (LR) in Arabic TC. Experiments are conducted on Aljazeera Arabic News (Alj-News) dataset. Experimental results recorded a precision of 96.5% on one category and above 90% for 3 categories out of the five categories of the Alj-News dataset. Regarding the overall performance, LR has recorded a macro-average precision of 87%, recall of 86.33% and Fmeasure of 86.5%.

Leen Al Qadi et al. [17] constructed a new dataset which contains almost 90k Arabic news articles with their tags from Arabic news portals, some of classical supervised machine learning classifiers were used. Namely: Logistic Regression, Nearest Centroid, Decision Tree (DT), Support Vector Machines (SVM), K-nearest neighbors (KNN), XGBoost Classifier, Random Forest Classifier, Multinomial Classifier, Ada-Boost Classifier, and Multi-Layer Perceptron (MLP). In pursuit of high accuracy, an ensemble model to combine the best classifiers together in a majority-voting classifier was implemented. The experimental results showed solid performance with a minimum F1-score of 87.7%, achieved by Ada-Boost and a top performance of 97.9% achieved by SVM.

Mokhtar Ali Hasan and Mohammed Abdullah Hassan [18] provide experimental evaluations of six well-known classification models in classifying a large Arabic corpus. These models are Nave Bayes (NB), Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (DT), and Stochastic Gradient Descent (SGD). We used a corpus consisting of 111,728 Arabic documents. Three performance metrics were applied to evaluate the experimental results of each model. The experimental results show that the Logistic Regression model scores the highest weighted F1 score, followed by SGD, SVM, NB, Random Forest, and DT.

Ahmed Omar et al. [19] constructed a standard multi-label Arabic dataset using manual annotation and a semi supervised annotation technique that can be used for short text classification, sentiment analysis, and multilabel classification. Then, they evaluated the topics classification, sentiment analysis, and multilabel classification. Based on that evaluation we found a relationship between topics published in OSNs and hate speech. The experimental results validate the effectiveness of the proposed technique.

Marwan Al Omari et al. [20] proposes a logistic regression approach paired with term and inverse document frequency (TF*IDF) for Arabic sentiment classification on services' reviews in Lebanon country. Reviews are about public services, including hotels, restaurants, shops, and others. Collected manually from Google reviews and Zomato, which have reached to 3916 reviews. Experiments show three core findings: 1) The classifier is confident when used to predict positive reviews. 2) The model is biased on predicting reviews with negative sentiment. Finally, the low percentage of negative reviews in the corpus contributes to the diffidence of logistic regression model.

Table 1 summarizes the previous papers that used logistic regression and different classification models on the Arabic language.

Table 1. Summarizing of Previous Papers

work	Dataset size	Feature vectorization methods	Experimental Results	LR Performance
Tarik Sabri et al.[15]	22429+5070 Docs	TF-IDF- Word count - Word2vec	F1-Score 73.42% to 98%	F1-SCORE 98.79
Mayy M. Al-Tahrawi [16]	1500 Docs	Normalized frequency	F1 Score 72.9% to 93.1%	F1-SCORE 86.5%
Leen Al Qadi et al. [17]	89189 Articles	TF-IDF	F1-SCORE 87.7% to 97.9%	F1-SCORE 97.9%
Mokhtar Ali Hasan and Mohammed Abdullah Hassan [18]	111728 Docs	BoW	Aver. F1-SCORE 89% to 96%	F1-SCORE 96%
Ahmed Omar et al. [19]	44000 Articles	BoW - TF-IDF - N-gram	Acc. 35.9% to 97.92%	Acc. 96.6%
Marwan Al Omari et al. [20]	3916 Reviews	TF-IDF	-	Macro-Precision 84%

The difference between this paper and previous studies is data size and feature vectorization methods used to evaluate the performance of using logistic regression classifier.

3. Experiment

This section presents an empirical study of three feature vectorization methods: BoW, N-grams and TF-IDF. A large data corpse (111728 documents) used to study, compare and evaluate these methods. Once feature vectorizations are calculated, logistic regression as a machine learning algorithm will be used. The results are established on the basis of the three statistical formulas which are precision, recall and F-measure.

3.1. Dataset

The dataset was collected by [14] and produced using a semi-automatic web crawling approach with 111,728 documents and 319,254,124 words found in three Arabic online newspapers: Assabah, Hespress, and Akhbarona. Five categories have been used to group the papers in the dataset: sport, politics, culture, economy, and diverse. It differs how many documents and words in each class. Table2 shows the number of documents in each category:

Table 2. Documents Distribution Among Categories

Category	*Number of Docs*
Culture	13738
Economy	14235
Diverse	16728
Politic	20505
Sport	46522
Total	*111728*

3.2. Pre-processing

Enhancing the performance of the models in natural language processing is largely dependent on data cleaning[21].So, In this experiment performing the following tasks as a text pre-processing of the studied dataset:

- Remove non-arabic words.
- Remove non-arabic letters: numbers, punctuation and symbols (* % $ = + …).

• Remove all diacritics of the Arabic words.

• Remove all Arabic stop words such as (‘عن’,’about’) , (‘ألى’,’to’)

• Normalize some letters that have different shapes in the same word. Some of the characters that has been normalized shown in Table 3.

Table 3. Some of the Characters has Normalized.

Character(s)	Normalized Form
إأآ	ا
ي	ى
ؤئ	ء
ة	ه

3.3. Feature vectorization

A method for converting text into numerical feature vectors that can be applied to machine learning is called feature vectorization[22]. BoW, TF-IDF, and N-gram are the three feature vectorization techniques has been used in this work.

- Bag of Words (BoW): Based on the frequency of each token's occurrence in text documents, Bag of Words (BoW) is a numerical representation of a text that turns the text into a numerical vector. Word order is lost and syntactic patterns are disregarded in BoW[23].
- Term Frequency (TF): The number of times a specific token appears in a document is known as its term frequency (TF)[24].
- Inverse Document Frequency (IDF): The score of the tokens across all documents is called the Inverse Document Frequency (IDF)[24].
- Term frequency–Inverse document frequency (TF–IDF) yields by Multiplying TF by IDF[25].
- N-gram: A set of n-tokens that appear in that sequence in a text set is known as an N-gram representation. It is more accurate at capturing both thematic and grammatical information. For instance, if n = 2, each phrase is formed from a succession of two words[26].

3.4. Performance metrics

To determine whether the classification model can accurately classify unseen data into the appropriate classes, the classifier's performance needs to be assessed after the algorithm has been chosen and constructed. Variety of techniques to assess the effectiveness of the classification algorithm has been applied, including the following definitions of f1-score, accuracy, precision, and recall:

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$F1 = \frac{2\times Precision \times Recall}{Precision + Recall} \quad (4)$$

where True Positive, True Negative, False Positive, and False Negative are denoted by the letters TP, TN, FP, and FN, respectively. The above metrics are combined with the micro-average measures in the multilabel categorization[19].

4. Results and Discussion

Using a random split technique, the dataset is tested with 80% of the data used for training and 20% for testing. All of the techniques in this study were evaluated on a 64-bit Windows 10 touring PC with a 32-GB RAM and a 512-GB SSD hard drive.

A number of feature vectorization methods (BoW, TF-IDF, N-gram (1,2), N-gram (1,3), and N-gram (1,4)) were to be compared. However, some techniques, like N-gram (1,3) and N-gram (1,4), need a lot of memory and processing power, thus our computers are unable to train them using this large dataset.

Table 4 illustrates the evaluation findings by F1 metrics, precision, and recall for each category. Out of all feature vectorization methods, the sport category gets the highest score, in contrast, the Economy category in the BOW approach has the lowest F1 score. The fact that there are substantially more documents in the sports category than in the economic category Table 2 implies the models may be biased in favor of the sports category's advantages. Nevertheless, the results were close, with the LR classifier achieving the best result of 96% using TF-IDF and N-gram (1,2) and BoW coming in second with 95% as shown in Table 5.

Table 4. Performance Evaluation of LR Classifiers Using BoW, TF-IDF, N-gram(1,2)

Category	*BoW*			*TF-IDF*			*N-gram (1,2)*		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
culture	0.96	0.95	0.95	0.94	0.95	0.95	0.96	0.95	0.96
diverse	0.96	0.97	0.97	0.95	0.97	0.96	0.96	0.98	0.97
economy	0.89	0.90	0.90	0.90	0.92	0.91	0.91	0.91	0.91
politic	0.91	0.91	0.91	0.93	0.91	0.92	0.92	0.92	0.92
sport	0.99	0.99	0.99	1.0	0.98	0.99	0.99	0.99	0.99
Average	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.95

Table 5. Accuracy Comparison Between Feature Vectorazation Methods

Feature Vectorization Methods	Accuracy
BoW	0.95
TF-IDF	0.96
N-gram (1,2)	0.96

5. Conclusions and Future Work

In this work, Evaluating the performances of various feature vectorization techniques utilizing large Arabic datasets comprising 111,728 texts using a logistic regression classifier. The three feature vectorization techniques used are BoW, TF-IDF, and N-gram. Removing stop words and normalized specific Arabic letters from the datasets before preprocessing them. The outcomes of these tests demonstrate that LR classifiers using TF-IDF and N-gram (1,2) performed the best, with an accuracy of 96% and BoW in second place with 95%.

The concept can be expanded to evaluate the efficacy of several classifiers by combining feature vectorization methods and comparing various classification models.

REFERENCES

[1] K. Sundus, F. Al-Haj, and B. Hammo, “A Deep Learning Approach for Arabic Text Classification,” in 2019 *2nd International Conference on new Trends in Computing Sciences* (*ICTCS*), Amman, Jordan: IEEE, Oct. 2019, pp. 1–7. doi: 10.1109/ICTCS.2019.8923083.

[2] H. El Rifai, L. Al Qadi, and A. Elnagar, “Arabic text classification: the need for multi-labeling systems,” *Neural Comput. Appl*., vol. 34, no. 2, pp. 1135–1159, Jan. 2022, doi: 10.1007/s00521-021-06390-z.

[3] A. Elnagar and O. Einea, “BRAD 1.0: Book reviews in Arabic dataset,” in 2016 IEEE/ACS *13th International Conference of Computer Systems and Applications* (*AICCSA*), Agadir, Morocco: IEEE, Nov. 2016, pp. 1–8. doi: 10.1109/AICCSA.2016.7945800.

[4] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, “Word Embeddings and Convolutional Neural Network for Arabic Sentiment Classification”. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2418-2427).

[5] A. I. Alharbi, P. Smith, and M. Lee, “Enhancing Contextualised Language Models with Static Character and Word Embeddings for Emotional Intensity and Sentiment Strength Detection in Arabic Tweets,” *Procedia Comput. Sci.*, vol. 189, pp. 258–265, 2021, doi: 10.1016/j.procs.2021.05.089.

[6] A. Elnagar, Y. S. Khalifa, and A. Einea, “Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications,” in *Intelligent Natural Language Processing: Trends and Applications*, vol. 740, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds., in *Studies in Computational Intelligence*, vol. 740. , Cham: Springer International Publishing, 2018, pp. 35–52. doi: 10.1007/978-3-319-67056-0_3.

[7] A. Elnagar, L. Lulu, and O. Einea, “An Annotated Huge Dataset for Standard and Colloquial Arabic Reviews for Subjective Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 142, pp. 182–189, 2018, doi: 10.1016/j.procs.2018.10.474.

[8] A. Al-Alwani and M. Beseiso, “Arabic Spam filtering using Bayesian Model,” *Int. J. Comput. Appl.*, vol. 79, no. 7, pp. 11–14, Oct. 2013, doi: 10.5120/13752-1582.

[9] Y. Li, X. Nie, and R. Huang, “Web spam classification method based on deep belief networks,” *Expert Syst. Appl.*, vol. 96, pp. 261–270, Apr. 2018, doi: 10.1016/j.eswa.2017.12.016.

[10] L. Lulu and A. Elnagar, “Automatic Arabic Dialect Classification Using Deep Learning Models,” *Procedia Comput. Sci.*, vol. 142, pp. 262–269, 2018, doi: 10.1016/j.procs.2018.10.489.

[11] S. M. Alzanin, A. M. Azmi, and H. A. Aboalsamh, “Short text classification for Arabic social media tweets,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 6595–6604, Oct. 2022, doi: 10.1016/j.jksuci.2022.03.020.

[12] M. M. Al-Tahrawi and S. N. Al-Khatib, “Arabic text classification using Polynomial Networks,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 27, no. 4, pp. 437–449, Oct. 2015, doi: 10.1016/j.jksuci.2015.02.003.

[13] A. H. Ababneh, “Investigating the relevance of Arabic text classification datasets based on supervised learning,” *J. Electron. Sci. Technol.*, vol. 20, no. 2, p. 100160, Jun. 2022, doi: 10.1016/j.jnlest.2022.100160.

[14] M. Biniz, “DataSet for Arabic Classification.” *Mendeley Data*, V2, 2018. doi: 10.17632/V524P5DHPJ.2.

[15] T. Sabri, O. E. Beggar, and M. Kissi, “Comparative study of Arabic text classification using feature vectorization methods,” *Procedia Comput. Sci.*, vol. 198, pp. 269–275, 2022, doi: 10.1016/j.procs.2021.12.239.

[16] Computer Science Department, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan and M. M. Al-Tahrawi, “Arabic Text Categorization Using Logistic Regression,” *Int. J. Intell. Syst. Appl.*, vol. 7, no. 6, pp. 71–78, May 2015, doi: 10.5815/ijisa.2015.06.08.

[17] L. A. Qadi, H. E. Rifai, S. Obaid, and A. Elnagar, “Arabic Text Classification of News Articles Using Classical Supervised Classifiers,” in 2019 *2nd International Conference on new Trends in Computing Sciences* (*ICTCS*), Amman, Jordan: IEEE, Oct. 2019, pp. 1–6. doi: 10.1109/ICTCS.2019.8923073.

[18] M. A. H. Madhfar and M. A. H. Al-Hagery, “Arabic Text Classification: A Comparative Approach Using a Large Dataset,” in 2019 *International Conference on Computer and Information Sciences* (*ICCIS*), Sakaka, Saudi Arabia: IEEE, Apr. 2019, pp. 1–5. doi: 10.1109/ICCISci.2019.8716479.

[19] A. Omar, T. M. Mahmoud, T. Abd-El-Hafeez, and A. Mahfouz, “Multi-label Arabic text classification in Online Social Networks,” *Inf. Syst.*, vol. 100, p. 101785, Sep. 2021, doi: 10.1016/j.is.2021.101785.

[20] M. Al Omari, M. Al-Hajj, N. Hammami, and A. Sabra, “Sentiment Classifier: Logistic Regression for Arabic Services’ Reviews in Lebanon,” in 2019 *International Conference on Computer and Information Sciences* (*ICCIS*), Sakaka, Saudi Arabia: IEEE, Apr. 2019, pp. 1–5. doi: 10.1109/ICCISci.2019.8716394.

[21] A. Ayedh, G. Tan, K. Alwesabi, and H. Rajeh, “The Effect of Preprocessing on Arabic Document Categorization,” *Algorithms*, vol. 9, no. 2, p. 27, Apr. 2016, doi: 10.3390/a9020027.

[22] J. Gao, K. Liu, B. Wang, D. Wang, and X. Zhang, “Improving deep forest by ensemble pruning based on feature vectorization and quantum walks,” *Soft Comput.*, vol. 25, no. 3, pp. 2057–2068, Feb. 2021, doi: 10.1007/s00500-020-05274-z.

[23] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown, “Text Classification Algorithms: A Survey,” *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.

[24] A. K. Joshi, “Natural Language Processing,” *Science*, vol. 253, no. 5025, pp. 1242–1249, Sep. 1991, doi: 10.1126/science.253.5025.1242.

[25] B. V. Babu et al., Eds., Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, vol. 236. in Advances in Intelligent Systems and Computing, vol. 236. New Delhi: Springer India, 2014. doi: 10.1007/978-81-322-1602-5.

[26] M. Damashek, “Gauging Similarity with n -Grams: Language-Independent Categorization of Text,” *Science*, vol. 267, no. 5199, pp. 843–848, Feb. 1995, doi: 10.1126/science.267.5199.843.