Exploring the Impact of Data Distribution and Augmentation on Machine Learning Models for Arabic Text Classification

W. Benaouda

LCPTS, FGE
USTHB University

DOI: 10.5281/zenodo.10475361

***Abstract*— In this paper, we conduct an investigation on the performance of different machine learning models for text classification based on topics. For this purpose, we used the SANAD corpus, which's a collection of Arabic texts consisting of three datasets that were compiled from various news portals. Furthermore , we combined these three datasets to evaluate how each one performs individually compared to the combined dataset.**

Keywords— *Text classification, Machine learning, Natural language processing, SANAD corpus, Topic identification.*

1 Introduction

The increase in textual data on the Internet has grown exponentially over years, This abundance of data has made it quite convenient to gather volumes of information from websites.

Text classification is an important task for organizing textual documents, and it is used in various fields such as sentiment analysis, author classification, as well as in the medical field and several other domains.

Most of the research in the field of NLP is conducted in the English language, while the Arabic language is very rich, and there are almost 226 million users creating online content in Arabic (Al-Tahrawi 2015). Arabic is used as a native language in 22 countries, with more than 447 million people using it as their mother tongue (Nguyen 2021). There are also several challenges associated with working with the Arabic language. For example, a letter can have multiple styles depending on its position in a word. The letter (ح) is written as (حـ) when it is at the beginning of a word, like in (حلم) , and (ـحـ) when it is in the middle of a word, and (ح) when it is at the end of a word, as in (صباح) (Saad 2010).

Machine learning models have been used in several studies and have yielded satisfactory results. For example, one can quote the use of Support Vector Machine (SVM), Decision-Tree, Naïve Bayes (NB), K-Nearest Neighbours, and Logistic Regression that were employed in (Itani 2012).

Additionally, multilayer perceptron with Mahalanobis distance and linear regression were utilized in other works (Al-Sarem 2019) to investigate the effect of the training set size on authorship attribution for instance.

The rest of the sections are organized as follows: related works are presented in Section 2, the description of the three datasets is provided in Section 3. Section 4 outlines the steps used for preprocessing . Results and discussion are presented in Section 5. Finally, Section 6 includes the conclusion and future work.

2 Related work

Over the past decades, several researchers have been interested in the field of text classification, both in English and Arabic. Numerous previous works have explored various approaches and methodologies. These studies, conducted in diverse domains, emphasize the importance of developing effective models for text classification in the Arabic language.

For example, a study conducted by (Hassan 2018) demonstrated the effectiveness of SVM and KNN machine learning models for theme classification. Similarly, in (Al-Harbi 2008), the performance of decision tree and linear support vector machine was tested. They used a dataset consisting of seven categories and found that the decision tree outperformed SVM in terms of accuracy.

A dataset composed of 2700 documents equally spread across nine categories is used in (Hmeidi 2015), in which the authors used a dataset comprising multiple classes containing 2700 Arabic articles. Five standard classification methods were employed, and the results indicated that SVM outperforms the other models.

The SANAD corpus, used in our study, was introduced by (Einea 2019) as a special resource aimed at enriching the field of natural language processing (NLP), particularly for the Arabic language. It consists of three datasets extracted from three different news portals. Additionally, it is publicly accessible and free of charge.

Thus, in (Elnagar 2020), several deep learning models were compared to assess their performance on both the Sanad and Nadia corpora. The results indicate that, for the Sanad corpus, the convolutional-GRU model achieved a minimum accuracy of 91.18%, while the attention-GRU model demonstrated the best performance with 96.94%.

3. Dataset

SANAD is a textual Arabic dataset composed of seven categories, namely: Finance, Medical, Culture, Politics, Religion, Technology, and Sports. The texts are collected from three news portals: AlArabiya, AlKhaleej, and Akhbarona. The three datasets contain a total of 194,797 labeled articles. Additionally, AlKhaleej is a balanced dataset, with each category containing 6,500 articles. However, AlArabiya and Akhbarona are slightly unbalanced datasets.

Figure 2 depicts the distribution of articles across the three datasets, while Figure 3 illustrates the percentage of articles in each dataset relative to the total number of articles.

TABLE I. Distribution of articles by topics

Label	AlArabiya	Akhbarona	AlKhaleej	Combined dataset
Finance	30,076	9,280	6,500	45,856
Sports	23,058	15,377	6,500	44,935
Culture	5,619	6,746	6,500	18,865
Tech	4,411	12,199	6,500	23,110
Politics	4,368	13,979	6,500	24,847
Medical	3,715	12,947	6,500	23,162
Religion	7,522	0	6,500	14,022

Table 1 represents the distribution of each category in the three datasets, as well as the distribution of the combined dataset. To enhance the visualization of the distribution, a graphical representation has been created in Figure 1 that shows the number of article distribution across the three datasets in SANAD corpus.

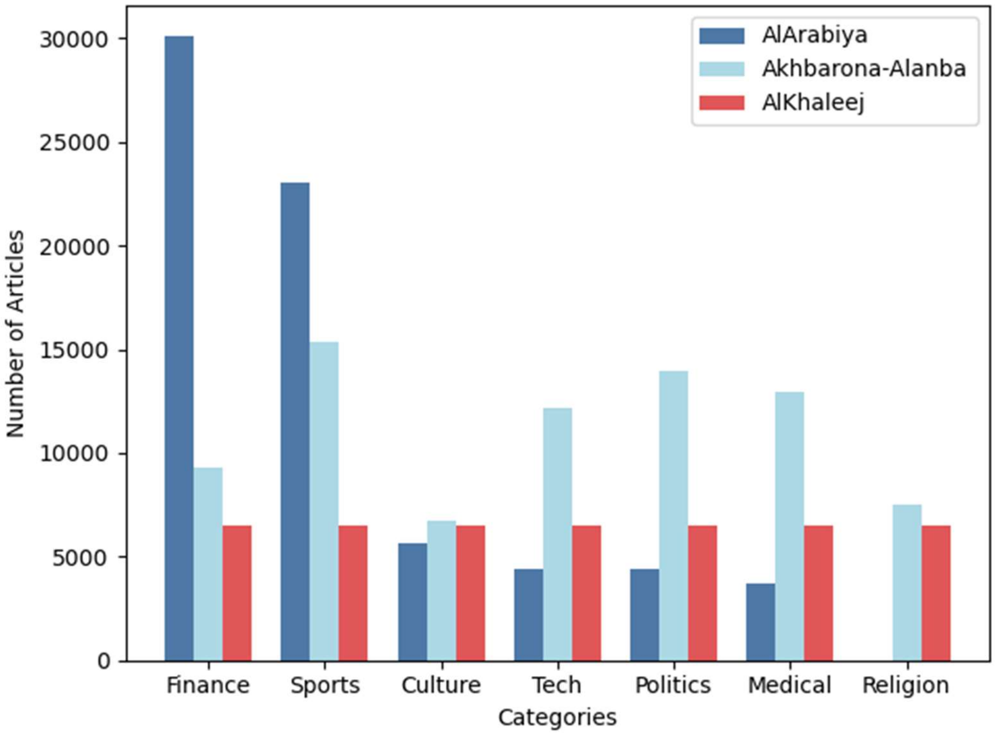

Fig. 1. Distribution of articles per label for each dataset.

4. Preprocessing

In this step, we removed diacritics, eliminated stopwords using the NLTK library, and performed normalization by replacing the letters "إ ,آ ,أ" with "ا" to simplify the text. Additionally, we removed punctuation marks and symbols.
Table 1 presents statistical information for the three datasets as well as the combined dataset.

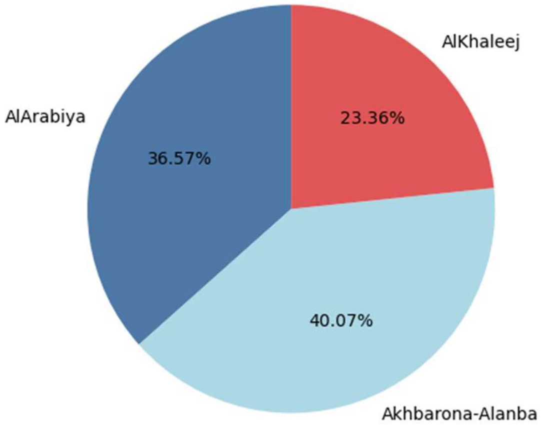

Fig. 2. Percentage Distribution of Each Dataset within the Combined Dataset.

TABLE II. Data statics

Dataset	AlArabiya	Akhbarona	AlKhaleej	Combined dataset
Total number of words	16,807,371	19,948,609	16,800,366	53,556,346
Number of unique words	670,181	1,049,102	905,531	1,924,060
Total number of characters	100,669,061	118,809,469	99,909,849	319,388,379

5. Results and discussion

In our experiment on the SANAD corpus, where 80% of the data was used for training and 20% was reserved for testing, we conducted a classification based on topics for the three datasets. Additionally, experiments were performed on the combination of datasets. For this purpose, we employed ten different machine learning models to assess the performances and evaluated these models using various metrics such as precision, recall, F1 score, and accuracy.
Tables 3, 4, 5, and 6 present the results for the different models used, while Figure 3 provides a graphical representation of the accuracy of each model for the various datasets.

TABLE III. Results of different machine learning models on the AlArabiya dataset.

Classifier	Precision	Recall	F1-score	Accuracy
Logistic Regression	96,92	96,65	96,78	98,09
Random Forest	96,54	93,14	94,75	96,89
Gradient Boosting	95,84	93,36	94,56	96,67
Multilayer Perceptron	96,87	96,02	96,43	97,89
Decision Tree	87,56	86,73	87,14	92,30
Ridge Classifier	96,75	95,92	96,32	97,79
Linear Support Vector Machine	**97,02**	**97,07**	**97,04**	**98,28**
AdaBoost	89,67	87,43	88,41	92,89
Bagging	92,64	90,21	91,38	94,85
Extra Trees	**97,02**	92,92	94,84	96,89

TABLE IV. Results of different machine learning models on Akhbarona dataset.

Classifier	Precision	Recall	F1-score	Accuracy
Logistic Regression	93.80	**93.34**	**93.55**	**94.02**
Random Forest	92.31	90.49	91.23	91.93
Gradient Boosting	91.19	90.25	90.67	91.29
Multilayer Perceptron	92.49	92.07	92.26	92.78
Decision Tree	83.30	83.27	83.28	84.34
Ridge Classifier	**93.81**	93.05	93.40	93.87
Linear Support Vector Machine	93.61	93.27	93.43	93.88
AdaBoost	82.79	80.05	80.91	82.35
Bagging	87.66	87.16	87.37	88.17
Extra Trees	92.93	90.97	91.77	92.41

TABLE V. Results of different machine learning models on the AlKhaleej dataset.

Classifier	Precision	Recall	F1-Score	Accuracy
Logistic Regression	97.66	97.66	97.66	97.65
Random Forest	96.91	96.84	96.84	96.82
Gradient Boosting	96.73	96.72	96.72	96.70
Multilayer Perceptron	97.78	97.79	97.78	97.77
Decision Tree	88.33	88.39	88.35	88.29
Ridge Classifier	97.60	97.60	97.60	97.58
Linear Support Vector Machine	**97.98**	**97.98**	**97.98**	**97.97**
AdaBoost	89.62	89.49	89.51	89.42
Bagging	92.34	92.29	92.28	92.24
Extra Trees	97.18	97.14	97.15	97.13

TABLE VI. Results of different machine learning models on the combined dataset.

Classifier	Precision	Recall	F1-Score	Accuracy
Logistic Regression	95.18	94.99	95.08	95.51
Random Forest	93.76	92.75	93.22	93.81
Gradient Boosting	92.81	91.93	92.35	92.92
Multilayer Perceptron	94.29	94.45	94.37	94.86
Decision Tree	84.37	84.20	84.29	85.77
Ridge Classifier	94.76	94.32	94.53	95.04
Linear Support Vector Machine	**95.21**	**95.02**	**95.11**	**95.53**
AdaBoost	83.59	82.40	82.75	83.92
Bagging	89.31	88.44	88.85	89.84
Extra Trees	94.27	93.22	93.71	94.27

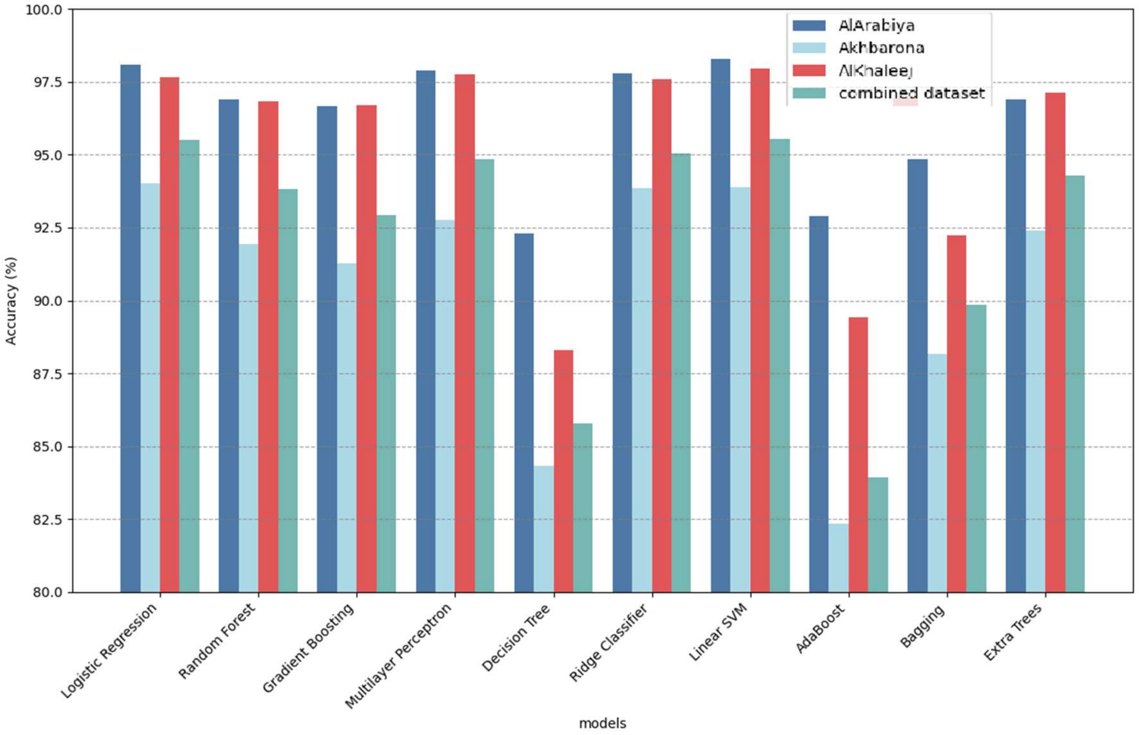

Fig. 3. Accuracy for Different machine learning Models by dfferent classifiers.

Figure 3 displays the accuracies obtained for the different cases. These results demonstrate that Linear Support Vector Machine has outperformed all other models in terms of accuracy. It

achieved an accuracy of 97.97% for the AlKhaleej dataset, also demonstrating better accuracies in all the other datasets: AlArabiya, Akhbarona, AlKhaleej, or the combined dataset.
On the other hand, the Decision Tree model was the least accurate classifier among the tested models, displaying an accuracy of 84.34% for the Akhbarona dataset. Although it produced acceptable results, as did all the other models.
Also, it is clear that the results for the combined dataset were inferior compared to the AlArabiya and AlKhaleej datasets, while they yielded superior results compared to the Akhbarona dataset. The maximum accuracy achieved by the Linear SVM on the combined dataset was 95.53.

6. Conclusion

In this investigation, various experiments on topic classification were conducted and analyzed. Firstly, an analysis and explanation of the SANAD corpus were carried out. In this study, an investigation was conducted on ten different machine learning models with four different datasets. The results indicate that the linear Support Vector Machine classifier surpassed all other models with the different datasets. On the other hand, the combined dataset yielded lower results compared to the two datasets named AlArabiya and AlKhaleej, and higher results than the Akhbarona dataset. This fact shows that having a large dataset is important for building a more efficient model, but the choice of a curated dataset is even more crucial.
In future work, we suggest exploring other techniques and models. Additionally, the use of stacking strategy is considered to enhance the performance of various machine learning models.

References

(Al-Harbi 2008) Sami Al-Harbi, Abdulrahman Almuhareb, Abdulmohsen Al-Thubaity, Mohammed Khorsheed, and Abdullah Al-Rajeh. 2008. Automatic Arabic text classification. In Proceedings of the 9th International Conference on Statistical Analysis of Textual Data. 77–83

(Al-Sarem 2019) Al-Sarem M, Emara A-H (2019) The effect of training set size in authorship attribution: application on short Arabic texts. Int J Electr Comput Eng 9(1):652–659

(Al-Tahrawi 2015) M.M. Al-Tahrawi, S.N. Al-Khatib, " Arabic text classification using polynomial networks ", journal king saud university 2015, Comput. Inf. Sci , Vol. 27 No. 4, pp 437–449

(Einea 2019) Einea, Omar, Ashraf Elnagar, and Ridhwan Al Debsi. "SANAD: Single-label arabic news articles dataset for automatic text categorization." *Data in brief* 25 (2019): 104076.

(Elnagar 2020) Elnagar, Ashraf, Ridhwan Al-Debsi, and Omar Einea. "Arabic text classification using deep learning models." *Information Processing & Management* 57.1 (2020): 102121.

(Nguyen 2021) Nguyen, T. N., Nguyen, N. P., Savaglio, C., Zhang, Y., & Dumba, B, " The Role of Artificial Intelligence (AI) in Healthcare Data Analytics ". JOURNAL ON ARTIFICIAL INTELLIGENCE TOOLS, Vol. 30 No. 06 N 08, December 2021 .

(Itani 2012) M. Itani, R. N. Zantout, L. Hamandi, and I. Elkabani, ''Classifying sentiment in arabic social networks: Naïve search versus Naïve bayes," in 2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), Dec. 2012, pp. 192–197

(Hassan 2018) Hassan, Geehan Sabah, " Categorization arabic text using svm and knn algorithms ". ." *International Journal of Engineering & Technology*, Vol. 7 No. 3.20, 2018, 906-909

(Hmeidi 2015) I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig, and N. A. Mahyoub, " Automatic Arabic text categorization: A comprehensive comparative study,". ," Journal of Information Science, Vol. 41 No. 1, 2015, pp 114-124.

(Saad 2010) M. K. Saad, W Ashour, " Arabic Morphological Tools forText Mining" In Int. Conf. Electr. Comput. Syst pp. 1–6, Northern Cyprus, 2010.