BERT-based models for classifying multi-dialect Arabic texts

Hassan Fouadi[1], Hicham El Moubtahij[2], Hicham Lamtougui[1], Ali Yahyaouy[1]
[1]Department of Computer Science, LISAC Laboratory, Faculty of Sciences Dhar EL Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco
[2]Higher School of Technology, Ibn Zohr University, Agadir, Morocco

Article Info

Article history:

Received Jul 14, 2023
Revised Dec 20, 2023
Accepted Jan 6, 2024

Keywords:

Arabic dialects
Arabic natural language processing
BERT models
Machine learning
Text classification

ABSTRACT

The area of natural language processing (NLP) is presently a rapidly developing field characterized by innovation and research. Despite this progress, several dialects of Arabic (DA) are classified as low-resource languages, making it challenging for NLP systems to process DA data. One approach to address this issue is to train NLP models on social media data sets containing DA texts. Therefore, these open-access social media datasets, as outlined in our paper, can serve as a valuable resource for developers and researchers involved in the processing of DA. To create our multilingual corpus, we gathered data from various datasets containing different versions of DA. These datasets will be used to classify texts in terms of sentiment classification, topic classification, and dialect identification. Our study contributes to the automated analysis of the classification of Arabic dialects. We aim to investigate and assess various machine learning and deep learning techniques, with a specific focus on utilizing the BERT model. The results of our experiments on our datasets show that DarijaBERT and DziriBERT trained on a similar DA outperform traditional machine learning methods and previous more general pre-trained models that were trained on multiple dialects or languages.

Corresponding Author:

Hassan Fouadi
Department of Computer Science, LISAC Laboratory, Faculty of Sciences Dhar EL Mahraz
Sidi Mohamed Ben Abdellah University
Route d'Imouzzer, BP 2626 Fez 30000 Fes - Fes, Morocco
Email: hassan.fouadi@usmba.ac.ma

1. INTRODUCTION

With more than 400 million speakers worldwide, Arabic, which has official status in 22 countries, has been as the fourth most widely used language in the world [1]. Furthermore, Arabic is known for its three main varieties, reflecting its rich history and diversity. Classical Arabic, which is used in the Qur'an, the holy book of Islam, and which is also used as a literary and formal language; modern standard Arabic (MSA), which is based on classical Arabic and is used in writing and formal speech throughout the Arabic-speaking world; and dialects of Arabic (DA), which is the spoken form of the language that varies from region to region and is used as the primary mode of communication both in written and spoken forms.

Over the past few decades, there have been efforts to improve the performance of natural language processing (NLP) systems for MSA, but these systems often struggle to perform well on DA data. This is in part because most DA, with the variety of dialects that exist, are considered low-resource languages, with a lack of labelled data available to build NLP systems, and a lack of dedicated models, particularly for its various dialects [2], [3]. Moreover, it is important to note that DA data from social media can be noisy and require careful processing. Several efforts aim to address these issues. The Gumar Corpus [4] (110M words) focusses on Gulf Arabic with sub-dialect annotations, while SUAR [5] (104,079 words) provides morphological annotations for the Saudi dialect. Another initiative [6] offers a dataset of over 50,000 tweets in various dialects for tasks like dialect detection and sentiment analysis. Additionally, research on classifying Arabic song lyrics [7] using machine learning (ML) and deep learning (DL) models achieved a 93% accuracy in binary dialect identification. Therefore, there is still a need for improvement in the categorization of DA text, particularly for contextual information and implicit meaning expressed in various practical contexts.

Thus, we focused this study on text classification which involves assigning relevant labels or categories to a textual document automatically. Recent advances in ML have contributed to a surge of interest in this field of research [8], leading to the development of successful text categorization systems that play a crucial role in applications such as sentiment analysis, topic classification, and dialect identification. To perform these text classification tasks, DL has been used as an effective approach which has been shown to outperform other ML methods in many cases. In recent years, among the DL models that have gained popularity for text classification tasks are transformer language models. These models have proven to be highly effective and have set new standards for accuracy in various downstream tasks.

An instance of a transformer language model is bidirectional encoder representations from transformers (BERT). Although the initial BERT was introduced in 2018, it was not until 2020 that specialized models for MSA, like AraBERT [9], emerged. The release of the first dialect-specific models in 2021, three years after the initial launch of the BERT model, illustrated the complexity and difficulties inherent in developing NLP systems designed for DA. The BERT model [10], [11] has been successful in achieving remarkable results in a wide range of NLP tasks and has gained extensive adoption within the field. The ability of the BERT model to transfer knowledge to downstream tasks makes it particularly useful for text classification, as it can be fine-tuned for specific datasets or tasks.

To overcome this challenge, we first begin by presenting the development of three categorization datasets specifically designed for DA: Senti-Dial, Identif_Dial, and Topic-Dial. These datasets were carefully curated from diverse news sources, encompassing a range of DA variations. Together, they form our multilingual corpus, which plays a crucial role in classifying written content in DA. Second, to improve the performance of NLP systems we pre-processed and cleaned the data before use. Finally, we investigated and evaluated the performance of four conventional ML techniques: stochastic gradient descent (SGD), logistic regression (LR), naïve Bayes (NB), and linear support vector classification (linear SVC), along with fine-tuning other three transformer-based language models: DarijaBERT, MARBERT, and DziriBERT. These were used to select and identify the most appropriate classifier for performing our tasks in DA. The current study shows that DarijaBERT and DziriBERT gave superior performance results in terms of accuracy, precision, recall, and f-measure compared to the MARBERT, SGD, LR, NB, and linear SVC classifiers.

The subsequent sections of this work are structured as follows: in section 2, we delineate the methodology employed in our study, covering aspects such as data collection, preprocessing, and the techniques applied. Section 4 is dedicated to presenting and discussing the results obtained. Finally, in section 5, we outline the conclusions drawn from our results.

2. THE PROPOSED APPROACH

This section outlines the ML techniques and pre-training models proposed for classifying DA texts, covering tasks like Arabic dialect identification, sentiment classification, and topic categorization. It emphasizes the prevalent use of pre-training models in recent years for their capability to identify features from raw data and their superior performance across diverse fields. The evaluation of specific NLP tasks employs various ML models and pre-training models, with the methodology architecture illustrated in Figure 1.

2.1. Machine learning algorithms

To enable a comprehensive comparison, we chose classifiers known for their strong performance in previous NLP tasks. These include LR, SGD, linear SVC, and NB [12]. The following sections will describe their implementation.

2.1.1. Logistic regression

LR is a supervised learning algorithm that is used for binary classification problems. It uses a logistic function (sigmoid function) to model the probability of an input belonging to a certain class. The LR function generates multiple linear functions that are expressed as (1).

$$\mathrm{Logit}(P) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \cdots + \beta_k Y_k \tag{1}$$

In the LR model, P represents the probability of the feature occurring, which is a crucial element in determining the likelihood of a particular outcome. Meanwhile, Y1, Y2, ..., and Yk represent the values of the predictor variables that influence this probability. The intercept of the model, denoted by β1, β2, ..., and βk, plays a fundamental role in adjusting the logistic function to best fit the data.

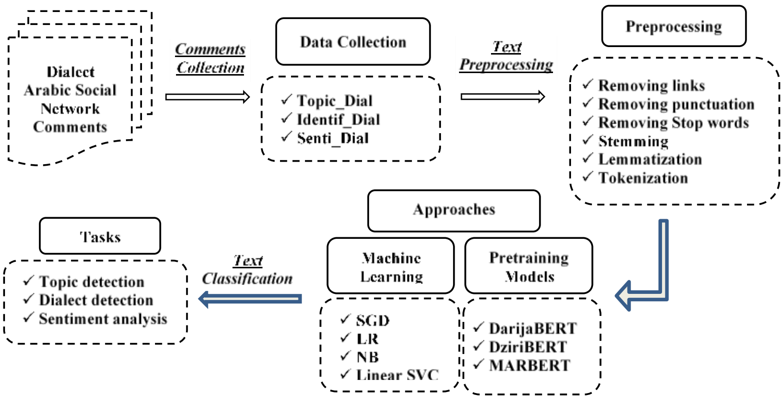


Figure 1. The general architecture of the proposed text classification system

2.1.2. Stochastic gradient descent classifier

SGD classifier is an optimization algorithm used to find the values of parameters that minimize the loss function and an iterative algorithm is used to find the optimal solution for the ML models such as linear regression and LR [13]. In each iteration, the algorithm updates the parameters by adjusting them in the direction of the negative gradient of the loss function concerning the parameters. One of the main advantages of the SGD classifier is that it can handle large datasets that are not possible to fit in memory.

2.1.3. Linear support vector classification

Linear SVC is a linear classification model. It is an extension of the support vector machine algorithm used for binary classification problems [14]. Linear SVC finds the best hyperplane that separates the data into two classes. The goal is to identify a hyperplane capable of effectively dividing the input space into two distinct classes. This hyperplane is defined by minimizing the loss function through the utilization of parameters w and b. The resulting equation representing the hyperplane is expressed as (2).

$$L(w,b) = w^t w + c \sum \max\left(0,1 - y^i \left(w^t F^{(i)} + b\right)\right)^2 \qquad (2)$$

2.1.4. Naive Bayes

The NB algorithm applied to text classification is a probabilistic method that assumes the independence of features given the class. Calculate the probability of a class given the text features using the as number (3). Where P(class) is the prior probability of the class, P(word|class) is the probability that a word occurs in the given class, and P(text): is the probability of the text [15].

$$P(class|text) = \frac{P(class) * \prod P(word|class)}{P(text)} \qquad (3)$$

2.2. BERT models

Recently, pre-trained word embedding techniques have been shown to achieve superior performance in various NLP tasks [16]. The primary concept underlying these models involves training them on extensive text data initially and subsequently fine-tuning them for specific NLP applications. This process enables the extraction of advantageous features and representations, which can enhance the models' performance. This approach is more effective than the existing systems.

2.2.1. BERT models for the Arabic language

BERT [10] is a highly utilized language modelling architecture that has demonstrated exceptional performance in various natural language understanding tasks, setting the standard in the field. The ability of BERT to understand the context of words by considering the words that come before and after them in a sentence, allows it to generalize to a wide range of tasks. As shown in Figure 2, fine-tuning the pre-trained

BERT model with a task-the specific layer enables BERT to adapt to the unique requirements of the task, resulting in improved performance compared to training from scratch.

In addition, there are multiple versions of BERT specifically designed for the Arabic language. These models have undergone training using extensive datasets of the Arabic language, allowing them to effectively capture the nuances and distinctive characteristics inherent in Arabic. It is crucial to note that the performance of a language model can be greatly influenced by the specific characteristics of the language itself. Using a model that has been trained in a similar language considerably improves its performance.

AraBERT [17], trained on a twenty-three Géga Byte text corpus of around three billion words, was the first model designed for MSA and was used as a reference until the release of ARBERT [18]. ARBERT has achieved the most advanced outcomes in subsequent tasks associated with MSA on a substantial corpus comprising sixty-one GB, containing 6.5 billion tokens. The authors of ARBERT also proposed a multidialectal model, called MARBERT [18]. Although multilingual models such as mBERT [19] are available for various languages, the scarcity of Arabic models that cater to a specific dialect still poses a considerable challenge.

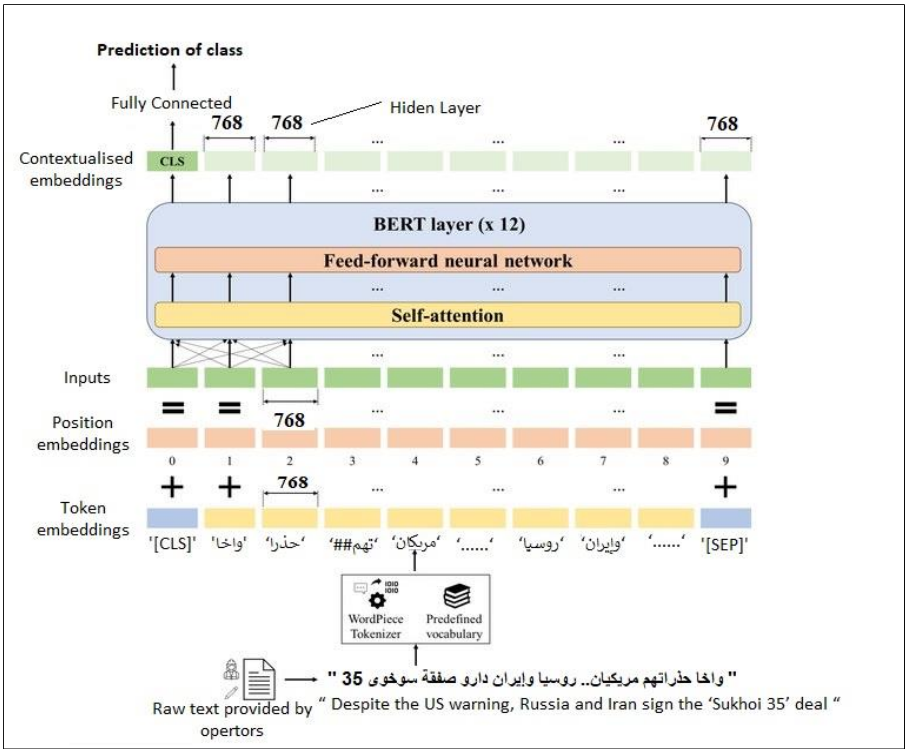

Prediction of class
Fully Connected
768
768
Hiden Layer
768
Contextualised embeddings
CLS
BERT layer (x 12)
Feed-forward neural network
Self-attention
Inputs
Position embeddings
768
0 1 2 3 4 5 6 7 8 9
768
Token embeddings
'[CLS]' 'واخا' 'حذرا' '##تهم' 'مريكان' '......' 'روسيا' 'وإيران' '......' '[SEP]'
WordPiece Tokenizer
Predefined vocabulary
" واخا حذراتهم مريكيان.. روسيا وإيران دارو صفقة سوخوى 35 "
Raw text provided by opertors
" Despite the US warning, Russia and Iran sign the 'Sukhoi 35' deal "

Figure 2. Example of a trained BERT for text classification

2.2.2. BERT models for dialect Arabic

Despite the existence of multilingual models for other languages, there is still a shortage of mono-dialectal models for Arabic. Only SudaBERT [20], TunBERT [21], DarijaBERT [6], and DziriBERT [22] exist for Sudan, Tunisia, Morocco, and Algeria among the twenty-two Arab countries. Diverse text data sources such as Twitter, public channels on Telegram, social media comments, and tweets were utilized to train these models. While supporting multiple dialects, the coverage of dialect-specific features and vocabulary is reduced. CAMeLBERT [23] provides three distinct models for MSA, DA, and classical Arabic. Furthermore, Qarib [24] is another model that combines MSA and multidialectal corpora. These models have undergone pre-training using a vast text corpus, enabling NLP practitioners to fine-tune them for specific tasks such as sentiment analysis, text classification, and named entity recognition.

2.3. Fine-tuned language models DA: MARBERT, DarijaBERT, and DziriBERT

Our specific tasks will utilize three DL methods that rely on fine-tuned language models, namely MARBERT, DarijaBERT, and DziriBERT. A comparison of the statistics of Arabic BERT models is presented. Specifically, Figure 3 displays the size of the vocabulary and the number of parameters, while Figure 4 provides information on the data size.

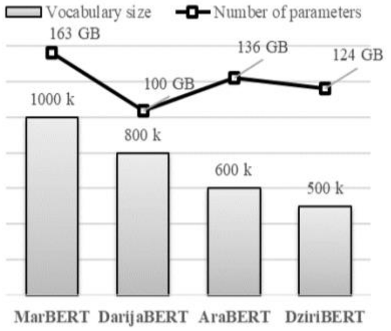


Figure 3. Comparison the size of the vocabulary and the number of parameters for different Arabic BERT models

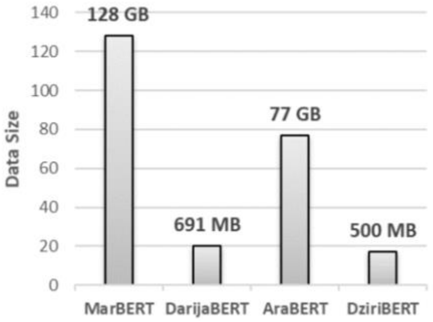


Figure 4. Comparison of data size for different Arabic BERT models

2.3.1. MARBERT

This is a language model trained specifically for Arabic and is capable of processing both DA and MSA. As there are multiple variations of Arabic, the MARBERT model was trained on a massive dataset consisting of 1 billion randomly sampled Arabic tweets. These tweets were selected from a much larger dataset of 6 billion tweets. The dataset used to train MARBERT consists of 15.6 billion tokens and amounts to a total of 128 GB of text [18].

2.3.2. DarijaBERT

This language model is specifically designed for Moroccan Darija, which is written in Arabic script, and created and provided by AIOX Labs, an AI company located in Rabat, Morocco. DarijaBERT has been trained using a variety of sources, including stories written in Darija, comments from 40 different Moroccan YouTube channels, and tweets that were collected using a specific keyword list. However, no written document was published explaining the specific details of the training process.

2.3.3. DziriBERT

DziriBERT is a BERT-based language model tailored to the Algerian Darja, which serves as the national vernacular of Algeria and shares significant mutual intelligibility with the Moroccan Darija [22]. It has been meticulously trained on a dataset comprising 1.2 million tweets gathered from prominent urban centres in Algeria. However, it is worth noting that in comparison to other language modelling training datasets, this corpus is relatively modest, as acknowledged by its developers.

3. DATASETS

The datasets used in this work were compiled from multiple sources and underwent a preprocessing step to clean and prepare the data. This included removing links, punctuation, and commonly used words (stopwords) to make the data more relevant and useful for analysis. As shown in Table 1, the study presents three main datasets that were created as part of this process. These datasets will be used to support the research, findings, and insights gained from the data analysis.

Table 1. Datasets distribution

Topic (Topic_Dial)		Identification (Identif_Dial)		Sentiment (Senti_Dial)	
Distribution	Numbers	Distribution	Numbers	Distribution	Numbers
Culture	02,050	Moroccan	10,987	Positive	30,000
Economy	01,800	Algerian	11,654	Negative	30,000
Sport	01,700	Tunisian	09,865	Neutral	30,000
Politics	01,885	Egyptian	08,454		
Diverse	02,420	Lebanese	11,654		
		MSA	10,643		

3.1. Presentation of data collection

3.1.1. Topic detection

The dataset comprises texts in the modern Arabic language texts extracted from newspaper articles. It includes words consisting of alphabets, numbers, and symbols. The presence of numeric and symbolic words in this dataset indicates the effectiveness and robustness of various Arabic text classification and document indexing techniques. The dataset consists of 9,854 sentences and 394,160 words structured in a CSV file and collected from 3 Arabic online newspapers: Ssabah, Hespress, and Akhbarona. According to Figure 5, the dataset consists of documents that have been divided into five distinct categories: culture, economy, sport, politics, and diverse.

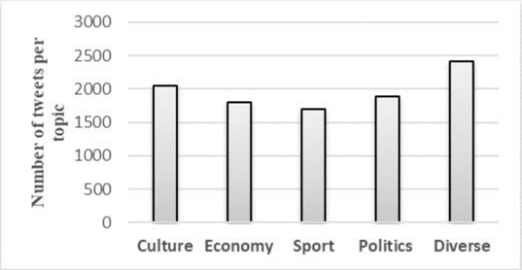


Figure 5. Topic distribution

3.1.2. Dialect detection

The total number of tweets is 63,257 with 6 different labels (MSA, Moroccan, Algerian, Tunisian, Egyptian, and Lebanese). It's a result of two data sets, the first was built by Mohamed VI Polytechnic University and the second from newspaper articles which contain only MSA language. Figure 6 presents the number of tweets per dialect.

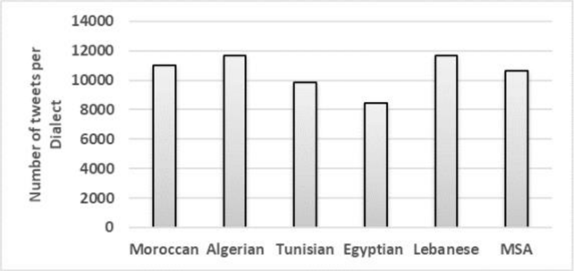


Figure 6. Dialects distribution

3.1.3. Sentiment analysis

The dataset comprises Arabic tweets classified into positive, negative, and neutral labels. The labels were obtained from the Arabic sentiment corpus created by Motaz Saad, which includes 32,000 tweets annotated with positive, negative, and neutral labels. By combining this corpus with another dataset created by Boujou, which includes 58,000 tweets, it is a balanced dataset with a total of 90,000 tweets. The files were transformed into a data frame using the R programming language before merging these datasets.

3.2. Data pre-processing steps

3.2.1. Noise removal

Given that our corpus comments are obtained through web scraping, they may probably contain unwanted elements from text data, such as punctuation and special characters. It also includes removing stop words, numeric characters, and non-textual data like URLs. Redundant or repeated words are identified and removed. These techniques improve data quality for accurate classification objectives and processing tasks in NLP.

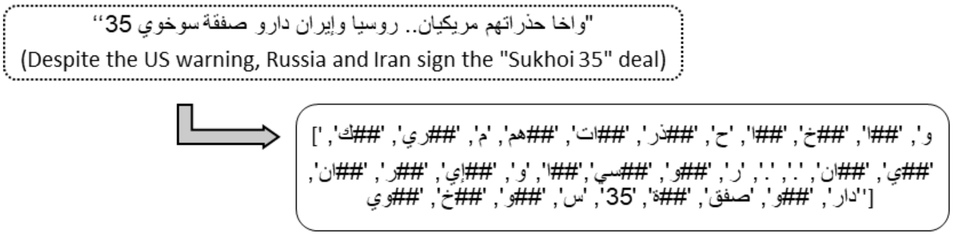
3.2.1. Stemming/lemmatization

Stemming is a crucial method when it comes to handling highly morphological languages like Arabic. In the stemming of Arabic words, the goal is to extract the stem or root of a word by eliminating all its prefixes and suffixes. The process involves removing the prefixes of the word starting from the right and moving towards the left, followed by the removal of any suffixes that may be present, also from right to left.

To eliminate the most encountered prefixes and suffixes, a lightweight stemmer algorithm was employed. The initial « و » and definite articles « لل », «وال », «ال », «كال », «فال », «بال » and suffixes « ها », «ان », «ات », «ون », «ين », «يه », «ية », «ه », «ة », «ي » are removed.

3.2.3. Tokenization

Tokenization refers to the procedure of dividing an input string of orthographic symbols into discrete "tokens," which can then be processed individually. Tokenisation systems are analytical tools that segment the text of a document into a sequence of tokens [25]. There is no single correct way to tokenise, and various options exist for specifying the splitting points. By default, non-letter characters are used to generate tokens consisting of a single word, which is ideal for text classification analysis. An example is shown in Figure 7.



"واخا حذراتهم مريكيان.. روسيا وإيران دارو صفقة سوخوي 35"

(Despite the US warning, Russia and Iran sign the "Sukhoi 35" deal)

['و', '##ا', '##خ', '##ا', 'ح', '##ذر', '##ات', '##هم', 'م', '##ري', '##ك',
'##ي', '##ان', '.', '.', 'ر', '##و', '##سي', '##ا', 'و', '##إي', '##ر', '##ان',
'دار', '##و', 'صفق', '##ة', '35', 'س', '##و', '##خ', '##وي']

Figure 7. Example of a tokenization text

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present a comprehensive overview of our experimental results and the diverse models used for the identification of Arabic dialects, the detection of sentiments, and the categorization of topics. The evaluation of these models involves calculating key performance metrics such as accuracy, macro-averaged precision, recall, and F1 scores across different datasets. To ensure a reliable and unbiased assessment, we rigorously followed a data-splitting protocol, dividing each dataset into three subsets: 70% for training, 15% for validation, and 15% for testing. This meticulous approach allowed for a thorough evaluation of the models' performance, with detailed discussions on experimental configurations, parameters, and results.

4.1. Evaluation metrics

The parameters help evaluate and assess the performance of supervised ML algorithms. From a classification point of view, terms such as 'true positive (TP)', 'true negative (TN)', 'false positive (FP)', and 'false negative (FN)' to make a comparison between the class labels [26]. In binary classification, a TP occurs when the classifier correctly identifies a positive instance, while a FP occurs when the classifier incorrectly labels a negative instance as positive. Similarly, a TN occurs when the classifier correctly identifies a negative instance, while a FN occurs when the classifier incorrectly labels a positive instance as negative.

In this study, we evaluated the performance of the classifier using four key metrics. Accuracy, a widely used measure in classification assessments, is calculated by dividing the number of correctly classified examples by the total number of examples. Precision, another crucial indicator, assesses the proportion of TP predictions among all positive classifier predictions. It's explicitly calculated as the ratio of TP to the sum of TP and FP. Recall, on the other hand, evaluates a classifier's capacity to identify positive examples by dividing the total number of correctly identified positive examples by the total number of positive examples in the dataset. Lastly, the F-measure, which balances precision and recall by computing their harmonic mean, is a valuable tool for optimizing system performance, allowing you to emphasize precision or recall based on your specific objectives.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{4}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{5}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{6}$$

$$F1 - score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{7}$$

4.2. Results

4.2.1. Dialect identification

The results for dialect identification are presented below, as depicted in Table 2, It's evident from the figure that BERT models outperform other models in terms of accuracy. The table shows a comparative analysis of different models, including SGD, LR, NB, linear SVC, DarijaBERT, MARBERT, and DziriBERT, based on various performance metrics. These metrics provide a comprehensive view of the models' performance across the dialect identification task, with BERT-based models, particularly DziriBERT and DarijaBERT, showcasing impressive accuracy and F1-score compared to other approaches.

Table 2. Results of dialect identification experiment

Model	SGD	LR	NB	Linear SVC	DarijaBERT	MARBERT	DziriBERT
Accuracy	0.72	0.72	0.75	0.76	0.83	0.81	0.87
F1-score	0.73	0.72	0.75	0.75	0.82	0.84	0.85
Precision	0.75	0.72	0.8	0.76	0.81	0.80	0.86
Recall	0.72	0.71	0.71	0.74	0.83	0.81	0.87

4.2.2. Topic classification

Table 3 provides a comprehensive summary of the research on topic classification. The table illustrates the performance of various models on this task. In particular, the BERT-based models, particularly DziriBERT, exhibit the highest accuracy, F-score, precision, and recall, surpassing the other models. These results highlight the effectiveness of BERT models in topic classification, and DziriBERT showcasing the most promising results in terms of accuracy and overall performance.

Table 3. Topic classification results

Model	LR	SGD	Linear SVC	DarijaBERT	MARBERT	DziriBERT
Accuracy	0.82	0.84	0.84	0.87	0.84	0.94
F1-score	0.82	0.72	0.84	0.88	0.86	0.93
Precision	0.59	0.65	0.65	0.86	0.83	0.92
Recall	0.59	0.45	0.45	0.89	0.81	0.96

4.2.3. Sentiment analysis

The performance of the sentiment analysis model implemented in BERT compared to the previous state-of-the-art systems is presented in Table 4. It was accurately found that the DziriBERT model produced the most favourable results. Specifically, this model achieved a score of 90% and an average F1-score of 89%.

Table 4. Sentiment analysis classification results

Model	SGD	LR	NB	Linear SVC	DarijaBERT	MARBERT	DziriBERT
Accuracy	0.77	0.76	0.73	0.76	0.89	0.80	0.90
F1-score	0.74	0.73	0.67	0.73	0.83	0.79	0.89
Precision	0.75	0.75	0.75	0.75	0.81	0.81	0.92
Recall	0.76	0.74	0.64	0.75	0.87	0.79	0.89

4.2.4. Discussions

Several studies have confirmed the effectiveness of a particular model in Arabic text classification. In our study, we investigated and evaluated the performance of several ML techniques including SGD, LR, NB, and linear SVC as well as three transformer-based language models: DarijaBERT, MARBERT, and DziriBERT to identify the most appropriate classifier for DA tasks. The experimental results were compared for each multilingual corpus. The findings indicated that the DziriBERT and DarijaBERT models outperformed other models for each dataset followed by MARBERT. In addition the traditional algorithms came last. However, it is important to note that better results were obtained for the topic classification.

The success of DarijaBERT and DziriBERT can be attributed to the fact that they are mono-dialectal models specifically trained on the Moroccan and Algerian dialects, respectively, known as Darija. This means that the model can focus on the unique linguistic characteristics of the dialect, allowing it to achieve higher accuracy on tasks related to Darija than more general models that are trained on multiple dialects or languages. So, using a model that has been trained in a similar language considerably improves its performance.

The use of models like MARBERT, trained across multiple dialects, presents a potential challenge. These models may struggle to accurately capture the nuances of individual dialects, particularly those with limited representation in their training data. Despite its broader training dataset, this limitation can lead to reduced performance in tasks specific to certain dialects. This highlights the significance of domain expertise and thoughtful model selection, emphasizing that success in dialect-related tasks involves not only large training data but also its quality and diversity, which significantly impact model performance.

5. CONCLUSION AND FUTURE WORKS

In this study, the DA data used were collected from different available datasets with different variants of AD. The datasets were preprocessed and cleaned before being used for modelling. The performance of four ML techniques was then tested and evaluated along with fine-tuning three transformer-based language models used to select and identify the most appropriate classifier to perform sentiment classification, topic classification, and dialect identification of Arabic. Based on the experimental results, it can be concluded that both DarijaBERT and DziriBERT demonstrated superior performance compared to other classifiers in terms of precision, accuracy, F-measure, and recall. Overall, the success of DarijaBERT and DziriBERT highlights the importance of developing models that are tailored to specific dialects or languages, particularly in cases where the linguistic characteristics of the dialect are markedly different from other languages or dialects. By focusing on the unique features of a specific dialect, models like DarijaBERT and DziriBERT can achieve superior performance compared to more general models, even when the latter have access to larger training datasets.

REFERENCES

[1] A. Farghaly and K. Shaalan, “Arabic natural language processing: challenges and solutions,” *ACM Transactions on Asian Language Information Processing*, vol. 8, no. 4, 2009, doi: 10.1145/1644879.1644881.

[2] A. Dahou, M. A. Elaziz, J. Zhou, and S. Xiong, “Arabic sentiment classification using convolutional neural network and differential evolution algorithm,” *Computational Intelligence and Neuroscience*, vol. 2019, Feb. 2019, doi: 10.1155/2019/2537689.

[3] T. M. Omran, B. T. Sharef, C. Grosan, and Y. Li, “Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach,” *Data and Knowledge Engineering*, vol. 143, 2023, doi: 10.1016/j.datak.2022.102106.

[4] S. Khalifa, N. Habash, D. Abdulrahim, and S. Hassan, “A large scale corpus of gulf Arabic,” *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pp. 4282–4289, 2016.

[5] N. Al-Twairesh *et al.*, “SUAR: towards building a corpus for the saudi dialect,” *Procedia Computer Science*, vol. 142, pp. 72–82, 2018, doi: 10.1016/j.procs.2018.10.462.

[6] E. Boujou, H. Chataoui, A. El Mekki, S. Benjelloun, I. Chairi, and I. Berrada, “An open access NLP dataset for Arabic dialects : data collection, labeling, and model construction,” *ArXiv-Computer Science*, pp. 1-10, 2021.

[7] M. El-Haj, “Habibi - a multi dialect multi national Arabic song lyrics corpus,” *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pp. 1318–1326, 2020.

[8] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, “Feature selection using an improved chi-square for Arabic text classification,” *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, 2020, doi: 10.1016/j.jksuci.2018.05.010.

[9] W. Antoun, F. Baly, and H. Hajj, “AraBERT: transformer-based model for arabic language understanding,” *ArXiv-Computer Science*, pp. 1-7, 2020.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, 2019.

[11] A. Vasvani *et al.*, “Attention is all you need,” *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[12] N. Seman and N. A. Razmi, “Machine learning-based technique for big data sentiments extraction,” *IAES International Journal of Artificial Intelligence*, vol. 9, no. 3, pp. 473–479, 2020, doi: 10.11591/ijai.v9.i3.pp473-479.

[13] S. Diab, “Optimizing stochastic gradient descent in text classification based on fine-tuning hyper-parameters approach. a case study on automatic classification of global terrorist attacks,” *ArXiv-Computer Science*, pp. 1-6, 2019.

[14] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.

[15] I. Wickramasinghe and H. Kalutarage, “Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation,” *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, 2021, doi: 10.1007/s00500-020-05297-6.

[16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.

[17] H. El Moubtahij, H. Abdelali, and E. B. Tazi, “AraBERT transformer model for Arabic comments and reviews analysis,” *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 379–387, 2022, doi: 10.11591/ijai.v11.i1.pp379-387.

[18] M. Abdul-Mageed, A. R. Elmadany, and E. M. B. Nagoudi, “ARBERT & MARBERT: deep bidirectional transformers for Arabic,” *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, pp. 7088–7105, 2021, doi: 10.18653/v1/2021.acl-long.551.

[19] A. Conneau *et al.*, “Unsupervised cross-lingual representation learning at scale,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020, doi: 10.18653/v1/2020.acl-main.747.

[20] M. Elgezouli, K. N. Elmadani, and M. Saeed, “SudaBERT: a pre-trained encoder representation for Sudanese Arabic dialect,” *Proceedings of: 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering, ICCCEEE 2020*. IEEE, 2021, doi: 10.1109/ICCCEEE49695.2021.9429651.

[21] A. Messaoudi *et al.*, “TunBERT: Pretrained contextualized text representation for tunisian dialect,” *Communications in Computer and Information Science*, vol. 1589 CCIS. Springer International Publishing, pp. 278–290, 2022, doi: 10.1007/978-3-031-08277-1_23.

[22] A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, "DziriBERT: a pre-trained language model for the Algerian dialect," *ArXiv-Computer Science*, pp. 1-6, 2021.

[23] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," *WANLP 2021 - 6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop*, pp. 92–104, 2021.

[24] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training BERT on Arabic tweets: practical considerations," *ArXiv-Computer Science*, pp. 1-6, 2021.

[25] A. Pasha *et al.*, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pp. 1094–1101, 2014.

[26] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117–126, 2016, doi: 10.1016/j.eswa.2016.03.028.

BIOGRAPHIES OF AUTHORS

Hassan Fouadi received a master's degree in computer science from the High School of Technology at the University Sidi Mohamed Ben Abdellah, Fez, Morocco. where he is currently pursuing a Ph.D. degree. His research interests include sentiment analysis, machine learning, deep learning, and natural language processing. He can be contacted at email: Hassan.fouadi@usmba.ac.ma.

Hicham El Moubtahij is currently a Professor of Computer Science at the University of Ibn Zohr, Agadir, Morocco. He received his Ph.D. in Computer Science from the University of Sidi Mohamed Ben Abdellah, Fez, Morocco in 2017. He is now a member of the Systems and Technologies of Information Team at the High School of Technology at the University of Ibn Zohr, Agadir. His current research interests include machine learning, deep learning, Arabic handwriting recognition, text mining, and medical imagery. He can be contacted at email: h.elmoubtahij@uiz.ac.ma.

Hicham Lamtougui received a master's degree in computer science from the Faculty of Science Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco. where he is currently pursuing a Ph.D. degree. His research interests include machine learning, deep learning, Arabic handwriting recognition, and text mining. He can be contacted at email: hicham.lamtougui@usmba.ac.ma.

Ali Yahyaouy is a professor at the Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Morocco. He obtained his Ph.D. degree at the University of Technology of Belfort-Montbéliard, France and Sidi Mohamed Ben Abdellah University Morocco in 2010. He is a member of the LISAC laboratory in Fez. He has been the coordinator of the University Qualification "Software Engineering and Multimedia" since 2014 and of the International Francophone Master "Web Intelligence and Data Science" since 2017. His research activities mainly concern machine learning, multi-agent systems, smart grids, and ITS. He can be contacted at email: ali.yahyaouy@usmba.ac.ma.