Journal of Computer Science

Original Research Paper

Improvement of Moroccan Dialect Sentiment Analysis Using Arabic BERT-Based Models

Ghizlane Bourahouat, Manar Abourezq and Najima Daoudi

ITQAN Team, LyRICA Laboratory, School of Information Science, Morocco

Article history
Received: 15-10-2023
Revised: 28-10-2023
Accepted: 20-11-2023

Corresponding Author:
Ghizlane Bourahouat
ITQAN Team, LyRICA Laboratory, School of Information Science, Morocco
Email: ghizlane.bourahouat@esi.ac.ma

Abstract: This study addresses the crucial task of sentiment analysis in natural language processing, with a particular focus on Arabic, especially dialectal Arabic, which has been relatively understudied due to inherent challenges. Our approach centers on sentiment analysis in Moroccan Arabic, leveraging BERT models that are pre-trained in the Arabic language, namely AraBERT, QARIB, ALBERT, AraELECTRA, and CAMeLBERT. These models are integrated alongside deep learning and machine learning algorithms, including SVM and CNN, with additional fine-tuning of the pre-trained model. Furthermore, we examine the impact of data imbalance by evaluating the models on three distinct datasets: An unbalanced set, a balanced set obtained through under-sampling, and a balanced set created by combining the initial dataset with another unbalanced one. Notably, our proposed approach demonstrates impressive accuracy, achieving a notable 96% when employing the QARIB model even on imbalanced data. The novelty of this research lies in the integration of pre-trained Arabic BERT models for Moroccan sentiment analysis, as well as the exploration of their combined use with CNN and SVM algorithms. Furthermore, our findings reveal that employing BERT-based models yields superior results compared to their application in conjunction with CNN or SVM, marking a significant advancement in sentiment analysis for Moroccan Arabic. Our method's effectiveness is highlighted through a comparative analysis with state-of-the-art approaches, providing valuable insights that contribute to the advancement of sentiment analysis in Arabic dialects.

Keywords: ANLP, Embedding, Arabic, Transformer, Sentiment Analysis, CNN and SVM

Introduction

Individuals from all around the globe are increasingly interacting through social media. This interaction produces a large amount of data, which, once thoroughly examined, may be used by decision-makers to make the best possible decision. This is where sentiment analysis comes into play.

Sentiment Analysis (SA) falls within the domain of Natural Language Processing (NLP) and focuses on recognizing the emotion or opinion expressed within a text. Many users express themselves through various online resources. As a result, it is necessary to analyze user-generated data to monitor public opinion and support decision-making.

When addressing Sentiment Analysis (SA), numerous obstacles must be tackled, including informal writing styles and language-specific difficulties. Moreover, numerous words in different languages possess diverse meanings and orientations. Consequently, tools and resources are scarce for all languages (Wankhade *et al*., 2022). The Arabic language belongs to this category and as a result, researchers face difficulties when conducting Arabic Sentiment Analysis (ASA).

The Arabic language is the world's fifth most commonly spoken language, with a community of billions of people. It has a unique set of characteristics that present difficulties in NLP. The latter is carried out by following a series of steps, beginning with data collection and ending with the performed task, which is in our research SA, as illustrated in Fig. 1.

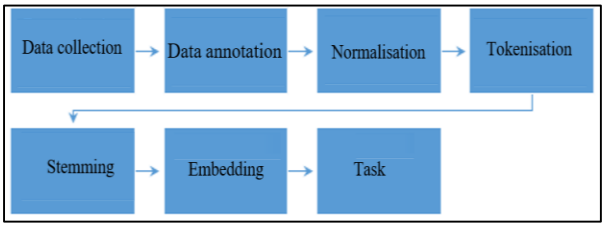

Fig. 1: ANLP process

Data collection and word embedding, which are two of the following steps in NLP, have a direct impact on the implemented model performance. As a result, we intend to test various combinations of Arabic BERT pre-trained models, namely AraBERT, ALBERT, AraELECTRA, QARIB, and CAMeLBERT, as both embedding layer and a baseline model, to investigate the impact of their structure. As data collection is also one of the investigated challenges, especially with the scarcity of datasets suited for ASA and the problem of data imbalance, to address the imbalance, we plan to employ two methods, namely, under-sampling and over-sampling. This involves enriching the unbalanced dataset by incorporating data from another unbalanced dataset, ultimately creating a newly balanced dataset.

This study introduces a novel approach to sentiment analysis by employing BERT models trained in the Arabic language for analyzing Moroccan sentiment, a domain that has seen limited exploration in prior research. Additionally, we extend the methodology by combining these BERT models with both CNN and SVM algorithms.

Context and Problem Statement

Moroccan Dialect Sentiment Analysis

Arabic Sentiment Analysis (ASA) faces distinct challenges attributed to the unique structure of the Arabic language, distinguishing it from languages such as English. Apart from its distinctive structure, the Arabic language includes three principal variations: Classical Arabic (CA), Standard Arabic (MSA), and Arabic Dialect (AD). CA is employed in religious and literary contexts, while MSA serves as the official language for education and formal communication. AD, on the other hand, lacks standardized spelling and predominantly represents the spoken form of Arabic, divided into five main groups: Egyptian, Levantine, Gulf, Iraqi, and Maghrebi (Ashi *et al*., 2019). Within the category of Maghrebi dialect, we specifically mention the Moroccan Dialect (MD), extensively used in everyday conversations and media within Morocco.

Research in the realm of Arabic Sentiment Analysis (ASA) is notably less extensive when compared to studies conducted in other languages. This disparity becomes particularly conspicuous when directing attention to the Moroccan Dialect (MD). ASA encounters increased complexity and challenges when dealing with dialects like MD (Tachicart *et al*., 2014). For instance, during the data collection process, available sources are not only limited in quantity but also frequently contain misspelled words and transliterations, particularly in the case of MD. Additionally, MD presents several other challenges, as identified by Adam (2019):

- No standard spelling pattern
- Sentences may contain words derived from SA, the Amazigh dialect "Tamazight", French, Spanish, and English
- Code-switching occurs when people switch from MD to another language or dialect in the same context

To tackle these challenges, researchers turned to ML and DL algorithms as they provide enhanced processing and analysis.

Data Collection

In the realm of NLP, data assumes a central and indispensable role. This applies to both rule-based methodologies, which make use of meticulously constructed lexicons and rules, and ML approaches, which depend on corpora and annotated datasets. While there exist numerous unannotated Arabic text corpora, the availability of morphological analyzers, Arabic lexicons and annotated corpora is limited. Furthermore, annotations beyond news and dialects are scarce. The significance of data availability becomes apparent in the NLP process depicted in Fig. 1, as data collection forms the foundation for subsequent work. Regrettably, the lack of high-quality and dependable Arabic resources remains one of the significant challenges, as reported by Ashi *et al*., 2019; Alayba and Palade, 2022; Almuzaini and Azmi (2020). Moreover, the data that is available presents the challenge of imbalanced data, which has a high impact on the developed model. The imbalanced data challenge can be resolved either by gathering more data, removing the redundancy, re-sampling the training dataset, or performing data augmentation. The scarcity of resources and data is aggravating when we are dealing with Arabic dialects, such as MD. For the latter, few valuable resources are available on open access.

Arabic NLP Techniques

In the field of NLP, various researchers use ML algorithms to deal with SA. Examined in studies by Hasan *et al*., 2018; Krishna *et al*. (2019), the approach encompassed the utilization of Naïve Bayes (NB), decision tree, and Support Vector Machine (SVM). Additionally, findings from (Yi and Liu, 2020)

accentuated that SVM is one of the most preferred methods employed for classification.

ML algorithms are widely employed in the field of NLP to tackle SA, as evidenced by studies conducted by Hasan *et al*., 2018; Krishna *et al*. (2019).

Moreover, within ASA, various ML methods have been employed. Nevertheless, three methods consistently demonstrate superior performance: SVM, NB, and K-Nearest Neighbor (K-NN). Notably in studies by Mohamed, 2022; Dehghani and Noughabi, 2022; Al sari *et al*. (2022) SVM consistently achieved the highest accuracy in ASA.

Moreover, SA has also capitalized on the advancements offered by DL algorithms. Studies conducted by Dang *et al*., 2020; Mulyo and Widyantoro, 2018; Cui *et al*., 2018; Sitaula and Shahi (2023) showcase the utilization of various DL algorithms, including CNN and LSTM, for SA. Usually, the CNN model demonstrated superior performance in SA. Moreover, CNN has exhibited effectiveness across a range of NLP tasks, consistently delivering noteworthy results in sentence classification, as illustrated in studies by Kalchbrenner *et al*., 2014; Kim (2014).

In ASA, DL algorithms are also implemented to resolve the faced challenges. Alayba and Palade (2022); Darwish *et al*. (2021); Guellil *et al*. (2021); Mohamed (2022), the use of the CNN model proved to be efficient since it achieved the best performance.

Transformers

The transformer architecture represents a ground-breaking development in the field of NLP by addressing sequence-to-sequence tasks and effectively handling long-range dependencies. The transformer exclusively utilizes self-attention to generate representations of both input and output sequences. Self-attention refers to an attention mechanism that connects various positions within a single sequence (Vaswani *et al*., 2017).

The Transformer has demonstrated remarkable success in numerous NLP tasks, including SA, with BERT serving as a prominent representative of the Transformer model.

BERT, an acronym for bidirectional encoder representations from transformers, is a language representation model that made its debut in late 2018. While BERT revolutionized the field, it's worth noting that several other pre-trained language models preceding BERT also employed bidirectional unsupervised learning, as mentioned by Hoang *et al*. (2019).

Building upon these insights, our approach aims to leverage multiple Arabic-BERT pre-trained models for SA. We will use the embeddings generated by these models and integrate them first with SVM, followed by integration with CNN. The choice of SVM and CNN is supported by their proven high performance in prior studies dedicated to ASA. By employing this combined approach, we aim to enhance the effectiveness and accuracy of sentiment analysis in the Arabic language, including the specific context of the Moroccan dialect.

The BERT model is a multi-layer bidirectional transformer encoder. It comes in two sizes: Base (12 encoders) and large (16 encoders) (24 encoders). To enable BERT to handle a wide range of downstream tasks, the input representation must be capable of representing both a single sentence and a pair of sentences unambiguously.

BERT offers a range of models designed for various languages, including Arabic. Various models exist such as AraBERT, ALBERT, AraELECTRA, QARIB, and CAMeLBERT. Given the focus of our research on the Arabic language, specifically the Moroccan dialect, our investigation will concentrate on exploring Arabic models that are based on BERT.

The objective of our work is to perform the task of SA in MD, especially by using Arabic transformer models. With this objective in mind, our research specifically concentrates on examining BERT models utilized for Sentiment Analysis (SA) in the context of the Moroccan or Maghrebi dialect. We aim to compare these models based on their effectiveness in accurately classifying sentiments within the Moroccan/Maghrebi dialect. According to our research findings, we notice that although several works dealt with performing the task of SA in the context of ANLP, few were concerned with the Maghrebi dialect in general and MD. In this section, we will present some work related to this field.

Darwish *et al*. (2021); Mhamed *et al*. (2021); Li *et al*. (2021) have used only ML/DL algorithms to perform SA on Maghrebi dialects. Mhamed *et al*. (2021) also performed the task of SA on the Algerian dialect by implementing the CNN model, which gave an accuracy of 89.5%.

Li *et al*. (2021), various ML models were implemented and compared such as SVM, NB, and logistic regression, this time to perform SA on MD. The best result was achieved by SVM combined with TF-IDF, with an accuracy of 84.31%. The majority of existing approaches for sentiment analysis in the Maghrebi dialect rely on traditional machine learning and deep learning techniques. However, there is limited research on the application of BERT models compared to other languages like English.

The existing research on Arabic Sentiment Analysis (ASA) utilizing Arabic-BERT pre-trained models, particularly for the Moroccan dialect, is relatively limited. However, notable findings have emerged from studies that combine AraBERT, QARIB, and CAMeLBERT with SVM, yielding interesting results as depicted in Table 1.

Table 1: Comparison of the techniques used in the related articles

Article	Technique	Arabic type	Performance %
Mhamed *et al*. (2021)	CNN	Algerian dialect dialect	Accuracy: 89.50
Li *et al*. (2021)	SVM combined with TF-IDF	Maghrebi dialect	Accuracy: 84.30
El Moubtahij *et al*. (2022)	AraBERT	Arabic dialect	Accuracy: 92.50
Mahdaouy *et al*. (2021)	AraBERT	Multi Arabic dialect	Accuracy: 89.60
Chowdhury *et al*. (2020)	AraBERT and SVM	MSA and Algerian dialect	Accuracy: 87.00
Farha and Magdy (2021)	QARIB	Multi Arabic dialect	Accuracy: 70.00
Abuzayed and Al-Khalifa (2021)	QARIB	Arabic dialect	F1-score: 75.00
Al-Yahya *et al*. (2021)	QARIB	Arabic dialect	Accuracy: 95.80
Mahdaouy *et al*. (2021)	QARIB	Jordanian dialectal and MSA	Accuracy: 92.00
Chowdhury *et al*. (2020)	QARIB and SVM	MSA and Algerian dialect	Accuracy: 90.00
Antoun *et al*. (2020)	AraELECTRA	Arabic dialect	F1-score: 57.20
Farha and Magdy (2021)	AraELECTRA	Arabic dialect	F1-score: 70.90
Kchaou *et al*. (2022)	ALBERT	Tunisian dialect	Accuracy: 67.10
Kchaou *et al*. (2022)	CAMeLBERT	Tunisian dialect	Accuracy: 43.00
Althobaiti (2022)	CAMeLBERT and SVM	Arabic dialect	F1-score: 45.10

In the study conducted by El Moubtahij *et al*. (2022), AraBERT was applied to a DA dataset for SA, achieving an accuracy of 92.5%. Furthermore, El Moubtahij *et al*. (2022) utilized AraBERT for the same task on a multi-dialect dataset, attaining an accuracy of 89.6%. Likewise, in Chowdhury *et al*. (2020), AraBERT, in conjunction with SVM, was utilized in a study involving a dataset comprising MSA and the Algerian dialect. The model achieved an accuracy of 87% in Sentiment Analysis (SA). QARIB, employed as both a classifier and an embedding model in a study (Farha and Magdy, 2021), attained a 70% accuracy when applied to a multi-dialect dataset. In another investigation (Abuzayed and Al-Khalifa, 2021), QARIB was employed for sentiment detection, yielding a 75% F1-score. QARIB's performance was further highlighted in various studies. In a study by Al-Yahya *et al*. (2021), QARIB served as an embedding model and classifier, achieving an impressive accuracy of 95.8%. Additionally, in a study conducted by Mahdaouy *et al*. (2021), QARIB demonstrated high accuracy, reaching 92%, when applied to a mixed corpus comprising Jordanian dialectal and MSA tweets. Noteworthy is its utilization by Chowdhury *et al*. (2020), where in conjunction with SVM, QARIB achieved an accuracy of 90% when applied to a combined dataset of MSA and Algerian dialects. AraELECTRA was implemented by Antoun *et al*. (2020) to perform ASA and achieved an F1-score of 57.20%. Farha and Magdy (2021), AraELECTRA was also used for ASA and the performance achieved was an F1-score of 70.9%. In turn, the ALBERT model was used by Kchaou *et al*. (2022) on a Tunisian dialect to perform SA and achieved an accuracy of 67.10%. Likewise, CAMeLBERT was implemented on the same dataset as mentioned for the ALBERT and obtained an F1-score of 43%. In (Althobaiti, 2022), CAMeLBERT was implemented together with the SVM classifier for the ASA task and achieved an F1-score equal to 45.10%.

Materials and Methods

Our approach is based on BERT models pre-trained on Arabic corpus, as shown in Fig. 2.

In our approach, we used the models AraBERT, QARIB, ALBERT, AraELECTRA, and CAMeLBERT to perform the task of SA, taking advantage of the whole architecture of the models. We also combined the aforementioned models with the algorithms SVM and CNN to see if they can improve the performance of the models. Furthermore, we carried out our tests on three datasets, which allowed us to explore the effect of data imbalance on the overall performance, as we will thoroughly explain in the present section.

The adopted approach is an improvement of our previous work (Bourahouat *et al*., 2022) by adding more embedding models and classifiers and testing our approach on various datasets.

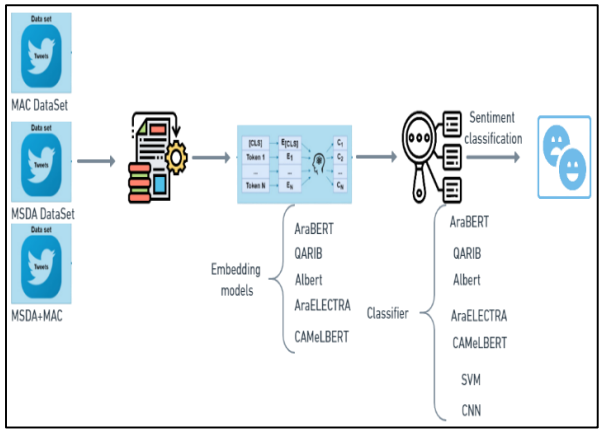

Fig. 2: Overall proposed approach

Dataset

The dataset plays a fundamental role in any research on Natural Language Processing (NLP). Selecting an appropriate dataset for our study posed a challenge, as we required a Moroccan Dialect (MD) dataset specifically tailored for Sentiment Analysis (SA) with an adequate amount of data. Our observation revealed a scarcity of open-source datasets available for MD. Consequently, in our research, we opted to work with two selected datasets to address this limitation: The Modelling Simulation and Data Analysis (MSDA) dataset (Boujou *et al*., 2021) and the Moroccan Arabic Corpus (MAC) dataset (Garouani and Kharroubi, 2021). We also combined the two to obtain a balanced dataset, the MSDA-MAC dataset:

- MSDA dataset**:** An open-access NLP dataset specifically designed for Arabic dialects, which was sourced from twitter. The dataset comprises over 50,000 tweets collected from five different national dialects, including the MD
- MAC dataset**:** A dataset that provides a free and large MD corpus consisting of 18.000 manually labeled tweets. The data contains 30,000 words labeled as positive, negative, and neutral
- MSDA-MAC dataset**:** This dataset was obtained by combining the MSDA and the MAC datasets. We tried to obtain a new balanced dataset with positive and negative classes. This method allowed us to create a new dataset based on the existing ones without having to go through web scraping

Arabic BERT-Based Models

Our research objective revolves around examining the impact of the structure of various Arabic pre-trained BERT models by exploring different combinations. To achieve this, we implemented several models including AraBERT, QARIB, ALBERT, AraELECTRA, and CAMeLBERT. Each model was utilized to perform Sentiment Analysis (SA), as they were trained on Arabic corpus, leveraging the comprehensive architecture of the respective model. By testing these combinations, we aim to gain insights into the effectiveness and performance of different Arabic pre-trained BERT models in the context of SA:

- AraBERT: AraBERT, derived from the BERT model, is a multi-layer bidirectional transformer. It is trained using MSA data, which restricts its suitability for tasks involving dialects
- QARiB: QARiB (QCRI Arabic and dialectal BERT) is a BERT model explicitly crafted for dialectal Arabic, akin to the focus of our study. In contrast to AraBERT, QARiB undergoes training on an extensive dataset, encompassing 420 million tweets and 180 million sentences of text. This extensive training data includes a wide range of dialectal Arabic content, enabling QARiB to effectively handle and understand the nuances of dialects in its language representations. (Abdelali *et al*., 2021)
- ALBERT: ALBERT is an optimized version of BERT, as indicated by its name. In comparison to BERT, the largest ALBERT model consists of approximately 70% of the parameters found in BERT-large (Lan *et al*., 2019)
- CAMeLBERT: CAMeLBERT constitutes an extensive ensemble of BERT models pre-trained on diverse Arabic texts as delineated in the study by Inoue *et al*. (2021). This compilation encompasses pre-trained language models specifically customized for MSA, DA, and CA
- AraELECTRA: AraELECTRA, which stands for Arabic ELECTRA, is a language representation model specifically designed for Arabic text (Antoun *et al*., 2020)

Given the adaptability of Arabic BERT-based models to various downstream tasks, we will leverage them for SA specifically targeting the Moroccan Dialect (MD). The central part of the architecture has been trained on extensive text corpora, resulting in frozen parameters within the internal layers. On the other hand, the outermost layers, responsible for task adaptation, will undergo fine-tuning. This approach entails utilizing the models for both embedding and classification purposes, as depicted in Fig. 3.

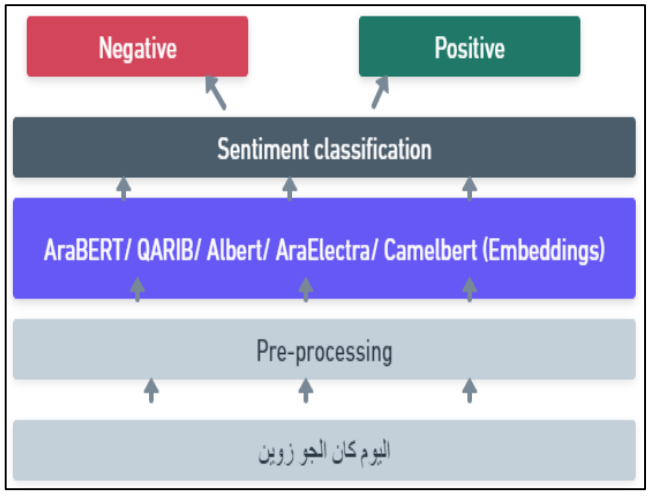

Fig. 3: Sentiment analysis using AraBERT, ALBERT, AraElecta, QARIB and CAMeLBERT

Embedding

During the encoding phase, we employ the embedding models of AraBERT, QARIB, ALBERT, AraELECTRA, and CAMeLBERT. The encode () function is employed to transform a list of strings into a corresponding list of vectors, effectively encapsulating the intents. Subsequently, we extract the pre-trained embeddings. The resulting output is a confidence level represented as a numerical value within the range of 0-1, where 1 signifies the utmost confidence level attainable. BERT provides contextual embeddings, signifying that the embedding of each word is influenced by its surrounding words. Nevertheless, given its contextual nature, we can infer that the initial token represented as '(CLS)', captures the overall context and can be considered as sentence embedding. This sentence embedding can then serve as input for the SVM algorithm (in the case of the ML approach), the CNN algorithm (for the DL approach), or the Arabic BERT-based models, as depicted in Fig. 4.

CNN Architecture

Our model employs a pre-trained Arabic-based model to embed sentences. These embeddings are then fed into a CNN model. The CNN model consists of multiple layers, starting with a convolution layer that utilizes various filters to extract features from different regions of the sentence. These filters convolve over the input, generating feature maps. Following this, the subsequent layer is the max-pooling layer, which identifies the most salient features within the produced feature maps. The flattened layer consolidates all the extracted features into a unified matrix. Ultimately, the fully connected layer, equipped with a sigmoid activation function, produces the output, delivering the classification of the input sentence as either positive or negative.

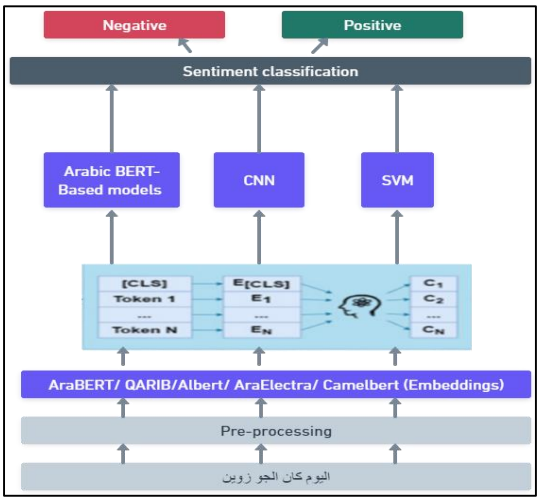

Fig. 4: Machine learning and deep learning approach based Arabic BERT models

Results and Discussion

In this section, we compare the performance of the different models we tested, namely using Arabic BERT-based models alone, combining them with the SVM algorithm, or combining them with the CNN algorithm. These tests we carried out on the MSDA dataset, the MAC dataset, and the MSDA-MAC dataset.

MSDA Dataset

As we've seen with Bourahouat *et al*. (2022), from the first analysis of the MSDA dataset, we discovered a significant challenge in our research due to the highly imbalanced nature of the dataset. This data imbalance poses a potential barrier to achieving optimal performance in our task.

To mitigate the impact of the data imbalance, we adopted a strategy of performing sentiment analysis with a focus on two classes: Negative and positive. To create a more balanced representation of these classes, we implemented an under-sampling technique.

As the dataset contains negative classes more than positive ones, we decided to apply an under-sampling of the dataset to get balanced data (6777 negative, 6777 positive) as shown in Fig. 5.

In our previous work Bourahouat *et al*. (2022), we implemented two Arabic BERT-based models, Ara-BERT and QARIB on the MSDA dataset. As a result, we found that QARiB performed best compared to AraBERT, giving an accuracy of 93 and 92% respectively. Therefore, we capitalized on the previous work and applied other combinations and models, as shown in Table 2.

We set the maximum sequence length at 160 as used in Bourahouat *et al*. (2022), the batch size used is 6 and the learning rate was fixed to 1e-6.

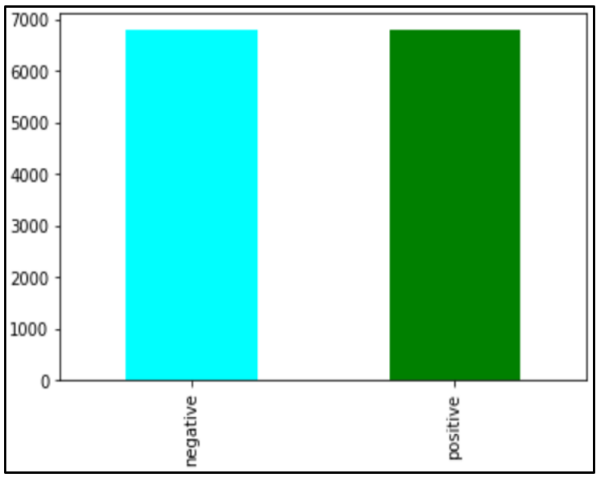

Fig. 5: Sentiment classes in the MSDA dataset

Table 2: Overview of the reviewed articles

Algorithm	Embedding	Epochs	Metric	Performance (%)
CNN	ALBERT	20	Accuracy	67
	AraELECTRA			72
	CAMeLBERT			73
	AraBERT			75
	QARiB			82
SVM	ALBERT	15	Accuracy	69
	AraELECTRA			71
	CAMeLBERT			74
	AraBERT			85
	QARiB			91
ALBERT	ALBERT	2	Accuracy	83
CAMeLBERT	CAMeLBERT			91
AraELECTRA	AraELECTRA			93
AraBERT	AraBERT			92
QARiB	QARiB			93

Table 3: Accuracy of the Arabic transformers applied to the MAC dataset

Algorithm	Embedding	Epochs	Metric	Performance (%)
CNN	ALBERT	15	Accuracy	62
	CAMeLBERT			65
	AraELECTRA			70
	AraBERT			76
	QARiB			79
SVM	ALBERT	15	Accuracy	71
	AraELECTRA			69
	CAMeLBERT			76
	AraBERT			85
	QARiB			89
ALBERT	ALBERT	2	Accuracy	74
CAMeLBERT	CAMeLBERT			75
AraELECTRA	AraELECTRA			93
AraBERT	AraBERT			93
QARiB	QARiB			95

From the results detailed in Table 1, we see that CAMeLBERT outperformed the other models when used by itself, resulting in an accuracy of 93%. However, compared with the results of (Bourahouat *et al*., 2022), the QARIB model and CAMeLBERT achieved the same accuracy of 93%.

MAC Dataset

The MAC dataset is highly imbalanced, as it consists of 9897 positive tweets and 3508 negative ones, as shown in Fig. 6.

For this dataset, we used 150 in the maximum length of the sequence, 6 in the batch size, and 1e-6c in the learning rate to get the results shown in Table 3.

For this dataset, QARIB achieved an accuracy of 95% when used by itself as an embedding model and a classifier, outperforming all the other models, whether they are used by themselves or combined with SVM or CNN. The results also showed that using QARIB as an embedding model and combining it with SVM also achieved the best performance of all the other models when combined with SVM, with an accuracy of 89%. The same conclusion was reached when combining QARIB with CNN, with an accuracy of 79%.

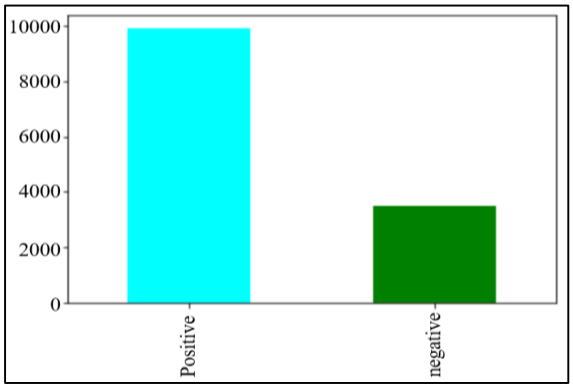

Fig. 6: Sentiment classes in the MAC dataset

MSDA-MAC Dataset

Although the MAC dataset is imbalanced, we achieved high performance with our approach, as detailed in the previous chapter. However, we wanted to see if applying our approach to a bigger balanced dataset would yield better results.

As we wanted to get a balanced dataset, we were faced with the following options:

- Under-sampling the MAC dataset to reduce the representation of the positive class, which would result in a balanced dataset but with less data
- Over-sampling the MAC dataset to increase the representation of the negative class, would result in a balanced dataset but with duplicated and/or synthetic data

To avoid losing in terms of data quality or data quantity, we decided to combine the two datasets, MSDA and MAC. As a baseline, we started with a dataset containing more data, which is the MAC dataset. We started then with this latter, to obtain an initial dataset of 13405 observations. Then we added 6 389 negative rows from the MSDA dataset since the latter has more positive observations than negative ones. In doing so, we obtained the MSDA-MAC dataset, which is a perfectly balanced dataset with real data where both classes, negative and positive, are equally represented. The MSDA-MAC dataset is 50% larger, with 9.894 observations in each class as illustrated in Fig. 7.

To utilize the pre-trained Arabic-based models, such as those mentioned in Table 4, it is necessary to specify the maximum length of the input sequence. In our particular case, we observed from Fig. 8 that the specific maximum length is less than 125 tokens. However, to ensure greater precision and accommodate potentially longer sequences, we chose to set the maximum length to 140 tokens. This decision allows us to capture more comprehensive contextual information and ensure that the input sequences are adequately accommodated within the models.

Table 4: Accuracy of the Arabic transformers applied to the MSDA-MAC dataset

Algorithm	Embedding	Epochs	Metric	Performance %
CNN	ALBERT	15	Accuracy	62
	CAMeLBERT			65
	AraELECTRA			70
	AraBERT			76
	QARiB			79
SVM	ALBERT	15	Accuracy	71
	AraELECTRA			69
	CAMeLBERT			76
	AraBERT			85
	QARiB			89
ALBERT	ALBERT	2	Accuracy	74
CAMeLBERT	CAMeLBERT			75
AraELECTRA	AraELECTRA			93
AraBERT	AraBERT			93
QARiB	QARiB			95

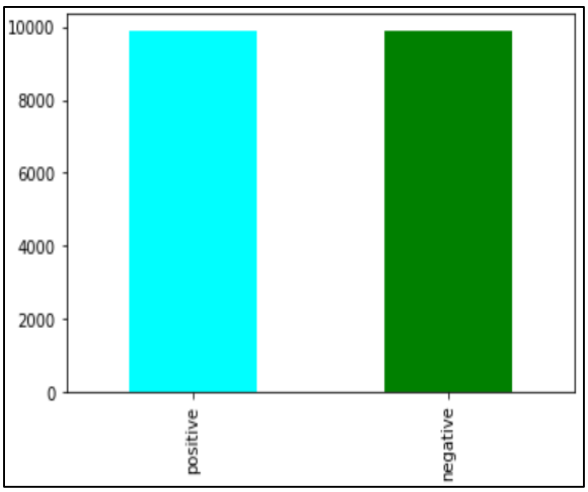

Fig. 7: Distribution of sentiment classes in the MSDA-MAC dataset

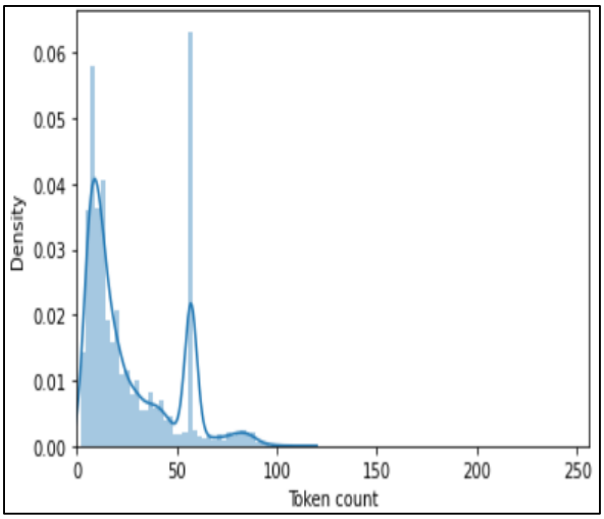

Fig. 8: Max length of tokens selected for the MSDA-MAC dataset

Once all the parameters are defined, we can implement the several models and algorithms as shown in Table 4.

Based on the results shown in Table 3, QARIB is still the best-performing model, whether it is used by itself or combined with SVM or CNN. Moreover, we noted that the overall accuracy has increased when using a balanced dataset, compared to the unbalanced datasets, with the highest accuracy reaching 96% for the MSDA-MAC dataset, for 95% for the MAC dataset and the MSDA dataset. This amelioration is even better for other models and combinations, reaching +14% for CAMeLBERT and +5% for AraELECTRA when combined with SVM.

The results presented in Tables 2-4 demonstrate that the utilization of transformer-based models, both as classifiers and embedding models, consistently outperforms traditional ML and DL algorithms such as SVM and CNN. This superior performance can be attributed to several factors, including the extensive language knowledge acquired by Transformers through training on diverse language modeling objectives. These findings highlight the significance of training language models on a wide range of topics and underscore the superiority of Transformers as embedding models for MD.

It is also worth mentioning that the ML algorithm SVM outperformed the DL algorithm CNN for all three datasets. Moreover, regarding the effect of imbalanced data, it is rather limited on already high-performing models, such as QARIB or AraBERT, but is more pronounced for other models, such as CAMeLBERT and AraELECTRA combined with SVM.

Our findings reveal a notable advancement in sentiment analysis performance when utilizing BERT-based models in contrast to their application with CNN or SVM. This discovery represents a significant contribution to the field, highlighting the efficacy of BERT models for improving sentiment analysis accuracy in Moroccan Arabic.

Conclusion

Sentiment analysis of the Arabic language remains a complex task and an ongoing challenge for researchers in the field of NLP. In this study, we have introduced our approach, which leverages multiple Arabic BERT-based models, including AraBERT, QARIB, ALBERT, AraELECTRA, and CAMeLBERT. These models were chosen to address the specific nuances and complexities of sentiment analysis in Arabic, aiming to improve the accuracy and effectiveness of sentiment classification in this language. In addition, we have investigated the impact of imbalanced data and detailed the main steps of our proposed approach. The process starts by feeding our model with pre-processed text sourced from the Moroccan dialect database. Subsequently, we employ BERT-based models for conducting sentiment analysis. Following this, we utilize the produced embeddings and categorize them. This categorization is achieved either

using the BERT-based models independently or by integrating them with the SVM and CNN algorithms, respectively. Moreover, our evaluation is based on several MD datasets, including a balanced dataset with an under-sampling technique, an imbalanced dataset, and a newly generated dataset from the previous ones. With well-tuned parameters of the QARIB model, we obtained the best accuracy in each of the three used datasets, reaching 96% for the MSDA-MAC dataset, 95% for the MAC dataset, and 93% for the MSDA dataset. QARIB is followed in terms of performance by AraBERT, AraELECTRA, CAMeLBERT, and ALBERT, respectively. These BERT-based models gave better results when implemented both as embedding models and classifiers than when combined with ML/DL algorithms such as SVM or CNN. Our work is still in progress. For future research directions, we aim to investigate other Arabic Transformers-based models for the task of SA, while also adapting our approach for other tasks.

Acknowledgment

We would like to thank all persons had contributed to this study. Also, we express our gratitude to the study's reviewers and editors.

Funding Information

This study did not receive funds from either public or private entities.

Author's Contributions

Ghizlane Bourahouat: Participated in all experiments, coordinated the data analysis, and contributed to the written of the manuscript.

Manar Abourezq: Contributed to the research planed and contributed to the written of the manuscript.

Najima Daoudi: Designed the research planed and organized the study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and that no ethical issues are involved.

References

Abdelali, A., Hassan, S., Mubarak, H., Darwish, K., & Samih, Y. (2021). Pre-training bert on arabic tweets: Practical considerations. *ArXiv Preprint ArXiv: 2102. 10684*. https://doi.org/10.48550/arXiv.2102.10684

Abuzayed, A., & Al-Khalifa, H. (2021). Sarcasm and sentiment detection in Arabic tweets using BERT-based models and data augmentation. *In Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 312-317. https://aclanthology.org/2021.wanlp-1.38

Adam, E. L. B. (2019). Sentiment Analysis for Moroccan Dialect. https://www.researchgate.net/publication/338237919_Sentiment_analysis_for_moroccan_dialect

Al Sari, B., Alkhaldi, R., Alsaffar, D., Alkhaldi, T., Almaymuni, H., Alnaim, N., ... & Olatunji, S. O. (2022). Sentiment analysis for cruises in Saudi Arabia on social media platforms using machine learning algorithms. *Journal of Big Data*, *9*(1), 21. https://doi.org/10.1186/s40537-022-00568-5

Alayba, A. M., & Palade, V. (2022). Leveraging Arabic sentiment classification using an enhanced CNN-LSTM approach and effective Arabic text preparation. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 9710-9722. https://doi.org/10.1016/j.jksuci.2021.12.004

Almuzaini, H. A., & Azmi, A. M. (2020). Impact of stemming and word embedding on deep learning-based Arabic text categorization. *IEEE Access*, *8*, 127913-127928. https://doi.org/10.1109/ACCESS.2020.3009217

Althobaiti, M. J. (2022). Bert-based approach to arabic hate speech and offensive language detection in twitter: Exploiting emojis and sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *13*(5). https://doi.org/10.14569/IJACSA.2022.01305109

Al-Yahya, M., Al-Khalifa, H., Al-Baity, H., AlSaeed, D., & Essam, A. (2021). Arabic fake news detection: Comparative study of neural networks and transformer-based approaches. *Complexity*, *2021*, 1-10. https://doi.org/10.1155/2021/5516945

Antoun, W., Baly, F., & Hajj, H. (2020). AraELECTRA: Pre-training text discriminators for Arabic language understanding. *ArXiv Preprint ArXiv:2012.15516*. https://doi.org/10.48550/arXiv.2012.15516

Ashi, M. M., Siddiqui, M. A., & Nadeem, F. (2019). Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets. *In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018* (4), 241-251. Springer International Publishing. https://doi.org/10.1007/978-3-319-99010-1_22

Boujou, E., Chataoui, H., Mekki, A. E., Benjelloun, S., Chairi, I., & Berrada, I. (2021). An open access nlp dataset for arabic dialects: Data collection, labeling and model construction. *ArXiv Preprint ArXiv: 2102. 11000*. https://doi.org/10.48550/arXiv.2102.11000

Bourahouat, G., Abourezq, M., & Daoudi, N. (2022). Leveraging moroccan Arabic sentiment analysis using arabert and qarib. In *The Proceedings of the International Conference on Smart City Applications*, 299-310. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-26852-6_29

Chowdhury, S. A., Abdelali, A., Darwish, K., Soon-Gyo, J., Salminen, J., & Jansen, B. J. (2020). Improving Arabic text categorization using transformer training diversification. *In Proceedings of the 5th Arabic Natural Language Processing Workshop*, 226-236. https://aclanthology.org/2020.wanlp-1.21

Cui, H., Lin, Y., & Utsuro, T. (2018). Sentiment analysis of tweets by CNN utilizing tweets with emoji as training data. *In Proc. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM)*, 1-8. https://sentic.net/wisdom2018cui.pdf

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483. https://doi.org/10.3390/electronics9030483

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., Bouamor, H., ... & Mubarak, H. (2021). A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, *64*(4), 72-81. https://doi.org/10.1145/3447735

Dehghani, M., & Noughabi, E. A. (2022). Sentiment analysis of persian political tweets using machine learning techniques. *Researchgate. Net*. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=U8Ebjo0AAAAJ&citation_for_view=U8Ebjo0AAAAJ:iH-uZ7U-co4C

El Moubtahij, H., Abdelali, H., & Tazi, E. B. (2022). AraBERT transformer model for Arabic comments and reviews analysis. *IAES Int. J. Artif. Intell*, *11*(1), 379-387. https://doi.org/10.11591/ijai.v11.i1.pp379-387

Farha, I. A., & Magdy, W. (2021). Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. *In Proceedings of the 6th Arabic Natural Language Processing Workshop* 21-31. https://aclanthology.org/2021.wanlp-1.3

Garouani, M., & Kharroubi, J. (2021). Mac: An open and free moroccan arabic corpus for sentiment analysis. *In The Proceedings of the International Conference on Smart City Applications*, 849-858. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-94191-8_68

Guellil, I., Adeel, A., Azouaou, F., Benali, F., Hachani, A. E., Dashtipour, K., ... & Hussain, A. (2021). A semi-supervised approach for sentiment analysis of arab (ic + izi) messages: Application to the algerian dialect. *SN Computer Science*, *2*, 1-18. https://doi.org/10.1007/s42979-021-00510-1

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11. https://doi.org/10.3390/mca23010011

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. *In Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 187-196. https://aclanthology.org/W19-6120

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size and task type in Arabic pre-trained language models. *ArXiv Preprint ArXiv:2103.06678*. https://doi.org/10.48550/arXiv.2103.06678

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *ArXiv Preprint ArXiv:1404.2188*. https://doi.org/10.3115/v1/P14-1062

Kchaou, S., Boujelbane, R., Fsih, E., & Belguith, L. H. (2022). Standardisation of dialect comments in social networks in view of sentiment analysis: Case of Tunisian dialect. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5436-5443. https://aclanthology.org/2022.lrec-1.582

Kim, Y. (2014). Convolutional neural networks for sentence classification. *ArXiv Preprint ArXiv:1408.5882*. https://doi.org/10.3115/v1/D14-1181

Krishna, A., Akhilesh, V., Aich, A., & Hegde, C. (2019). Sentiment analysis of restaurant reviews using machine learning techniques. *In Emerging Research in Electronics, Computer Science and Technology: Proceedings of International Conference, ICERECT 2018*, 687-696. Springer Singapore. https://doi.org/10.1007/978-981-13-5802-9_60

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *ArXiv Preprint ArXiv:1909.11942*. https://doi.org/10.48550/arXiv.1909.11942

Li, H., Ma, Y., Ma, Z., & Zhu, H. (2021). Weibo text sentiment analysis based on bert and deep learning. *Applied Sciences*, *11*(22), 10774. https://doi.org/10.3390/app112210774

Mahdaouy, A. E., Mekki, A. E., Essefar, K., Mamoun, N. E., Berrada, I., & Khoumsi, A. (2021). Deep multi-task model for sarcasm detection and sentiment analysis in Arabic language. *ArXiv Preprint ArXiv:2106.12488*. https://doi.org/10.48550/arXiv.2106.12488

Mhamed, M., Sutcliffe, R., Sun, X., Feng, J., Almekhlafi, E., & Retta, E. A. (2021). Improving Arabic sentiment analysis using CNN-based architectures and text preprocessing. *Computational Intelligence and Neuroscience*, *2021*. https://doi.org/10.1155/2021/5538791

Mohamed, A. (2022). SVM and naive bayes for sentiment analysis in Arabic. https://doi.org/10.21203/rs.3.rs-1631367/v1

Mulyo, B. M., & Widyantoro, D. H. (2018). Aspect-based sentiment analysis approach with CNN. In *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 142-147. IEEE. https://doi.org/10.1109/EECSI.2018.8752857

Sitaula, C., & Shahi, T. B. (2023). Multi-channel CNN to classify nepali COVID-19 related tweets using hybrid features. *Journal of Ambient Intelligence and Humanized Computing*, 1-10. https://doi.org/10.1007/s12652-023-04692-9

Tachicart, R., Bouzoubaa, K., & Jaafar, H. (2014). Building a Moroccan Dialect Electronic Dictionary (MDED). *In 5th International Conference on Arabic Language Processing* 216-221.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780. https://doi.org/10.1007/s10462-022-10144-1

Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. *Complex and Intelligent Systems*, *6*(3), 621-634. https://doi.org/10.1007/s40747-020-00155-2