Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.58, p.920-930. March/April, 2026.
DOI: https://doi.org/10.5935/jetia.v12i58.3205

RESEARCH ARTICLE OPEN ACCESS

FINE-TUNING STRATEGIES FOR SENTIMENT ANALYSIS IN THE ALGERIAN DIALECT: A COMPARATIVE STUDY ON DZIRIBERT

Salima Brachemi-Meftah[1], Fatiha Barigou[2]

[1,2]Laboratoire d'Informatique d'Oran Université Oran, Algeria.

[1]https://orcid.org/0009-0006-6622-0681, [2]https://orcid.org/0000-0001-5444-4000

Email: brachemi.meftah.s@gmail.com

ARTICLE INFO

Article History
Received: December 30, 2025
Reviewed: January 31, 2026
Accepted: March 10, 2026
Published: April 30, 2026

Keywords:
Sentiment analysis,
Arabic Algerian dialect,
DziriBERT,
Fine-tuning,
LoRA tuning.

ABSTRACT

The aim of this work is to explore sentiment analysis in the Algerian dialect through the adaptation of pre-trained linguistic models. We focus on DziriBERT, a model specifically developed for the Algerian dialect and derived from the BERT (Bidirectional Encoder Representations from Transformers) model, recognised for its performance in various natural language processing tasks. Three fine-tuning approaches were studied: full fine-tuning, freeze tuning, and LoRA tuning (Low-Rank Adaptation). Experiments were conducted on two separate corpora: ADArabic and the corpus introduced by Adouane, to evaluate the robustness and generalisation of DziriBERT model. The results show that the LoRA method achieved the best performances, on the ADArabic corpus, it achieved 83.26% in terms of accuracy and 80.94% for F1-score. On the Adouane corpus, LoRA reached the highest performance, with an accuracy of 85.28% and an F1-score of 82.52%. These results confirm the relevance of using DziriBERT for sentiment analysis in the Algerian dialect and highlight the effectiveness of LoRA tuning as a lightweight and efficient alternative to full fine-tuning, with significantly reducing the number of adjustable parameters.

I. INTRODUCTION

Sentiment analysis (SA) is regarded as one of the most active research areas in Natural Language Processing (NLP). Sentiment analysis aims to identify and classify the polarity, the emotions and opinions expressed in texts [1]. It has become a major area of research with the rise of social networks, with millions of users sharing their opinions on products, events or personalities on a daily life [2], [3], financial domain [4] and even in the health sector [5]. In this paper, we focus on sentiment analysis for the Algerian dialect. Currently, the Algerian dialect (AlgD) occupies an important place in the Maghreb region dialects and developing effective sentiment analysis methods for AlgD is crucial to better understand users' opinions on social media. AlgD is based on the Arabic language [6], [7], it is used extensively in everyday conversations, informal exchanges and online communication. It is characterised by a mixture of Arabic, French, Tamazight, and sometimes Spanish and Turkish. This linguistic diversity makes it particularly difficult to process automatically, as it has no standardized orthography and varies from one region of the country to another [8-10].

Initially, sentiment analysis was based on sentiment lexicon techniques [11]. Later, Artificial Intelligence took up the challenge through machine learning (ML) [12-15], or through deep learning (DL) as in [10], [16], [17]. Although DL is recognised for its remarkable performance in various NLP tasks, its effectiveness remains limited when datasets are small or when text representations are based on traditional approaches (such as binary representation). Indeed, in the study by [18] on SA for the Algerian dialect, the results obtained using traditional ML models such as Support Vector Machines (SVM) and Naive Bayes (NB) were superior to those produced by deep models such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The emergence of Transformers and large language models has improved the field of NLP by enabling models to understand context better, handle long sequences of text and achieve higher performance for a variety of NLP tasks. For example, BERT models [19] achieved state of the art results on numerous NLP tasks including sentiment analysis [19]. However, their performance depends on the data they were trained on which makes fine-tuning [20] essential for adapting them to specific tasks and different linguistic contexts. The Algerian dialect remains a particularly challenging case due to the lack of large annotated datasets, which makes it difficult for pre-trained models to work effectively.

This limitation is illustrated in the work of [10] where the performance of the BERT model proved to be inferior to that obtained using deep learning with CNN and Long Short-Term Memory (LSTM) and also traditional machine learning like SVM. This gap can be explained by the fact that the BERT model [19] was trained on a large English-language corpus and is therefore not adapted to the linguistic particularities of the Algerian dialect. Abdaoui et al. [21] developed DziriBERT, a model derived from BERT and refined from approximately 15 MB of texts. This model makes it possible to generate contextual representations adapted to AlgD. The pre-trained model DziriBERT was then refined to create the DziriBERT-Sentiment classifier, trained on the TWIFIL corpus [10].

The results obtained by DziriBERT-Sentiment surpassed those reported for the CNN, LSTM and BERT models used by previous studies [10] on the same corpus (TWIFIL). In this work, we investigate the DziriBERT-Sentiment model. We want to take advantage of fine-tuning techniques in order to adapt the model for use on other corpora. Additionally, we explore a novel fine-tuning method, LoRA tuning [22], which trains a reduced number of parameters thereby contributing to green AI applications. Furthermore, [23] showed that a moderately sized corpus (4 GB) could offer performance close to that obtained with a much larger corpus (130 GB). We therefore want to examine this finding in Algerian dialect sentiment analysis. This work is structured around the following research questions:

(I) To what extent does the DziriBERT-Sentiment classifier exhibit robustness when applied to new corpora?

(II) Is additional fine-tuning necessary to further improve the performance of the DziriBERT model for sentiment analysis in Algerian dialect when applied to new corpora?

(III) Which fine-tuning strategy (full fine-tuning, freeze tuning, or LoRA) achieves the most reliable performance gains?

(IV) How does the size of the training dataset affect the classifier's effectiveness, and what is the data volume required to obtain satisfactory performance levels?

The rest of this paper is organized as follows: Section 2 provides the background, offering a brief description of the key concepts discussed in this paper and review related works addressing sentiment analysis of AlgD. Section 3 presents our proposed method and data used in this research. Section 4 presents the whole set of results obtained. Finally, section 5 concludes the paper.

II. THEORETICAL REFERENCE

II.1 FUNDAMENTAL CONCEPTS

This section provides a brief overview of the important concepts underlying this work. These include sentiment analysis, the linguistic characteristics of the Algerian dialect, Transformer description, and strategies for fine-tuning pre-trained language models.

II.1.1 Sentiment Analysis

Sentiment analysis is one field of NLP. Its applications cover several sectors: marketing decision support, political intelligence, analysis of social trends and even the detection of hate speech [1]. SA is interested in studying the polarity of a text, whether it is negative or positive, this task is therefore a binary classification [24]. Other researches add a neutral class [10], [25], meaning that there are three classes to study.

II.1.2 Algerian Dialect

Algeria is one of 22 Arabic countries that speak Arabic. Modern Standard Arabic (MSA) is used in books, newspapers, official documents, etc. For oral communication in the daily life, people speak Arabic dialects. These dialects differ from country to other, from region to region in the same country. In Algeria, this dialect is called Algerian Dialect (AlgD). MSA is normalized and standardized but dialects are not. AlgD is characterized by: informal language, agglutinated negation, code switching (words from several languages in the same text), different writing script (Arabic, Latin or Arabizi) [6], [16], [26]. The Arabizi is recognised by the use of Latin letters and some numbers to designate some letters that do not exist in Latin. The letter "ع" in Arabic does not exist in Latin script, the internet users replace "ع" by a digit "3" ("ح" by "7", "ق" by "9", …).

II.1.3 BERT: Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) developed by Google Research in 2018, is a powerful NLP model. BERT was trained using Books Corpus (800 million words) and Wikipedia (2.5 billion words) [19], [27]. BERT is based on a transformer architecture, using self-attention mechanisms to comprehensively understand the contextual meaning of words in sentences. Two main phases are followed by the model: pre-training and fine-tuning [19], [20]:

1. Pre-training: BERT learns from unlabelled data.
2. Fine-tuning: adjusting BERT's pre-trained parameters using labelled data.

Given the success of transformers, other models have been developed for multi-languages mBERT [19] or dedicated to specific languages such as CamemBERT [23] for French language. For Arabic language, AraBERT [28] was pre-trained using 23 GB of MSA text extracted from Wikipedia.

II.1.4 Fine-Tuning

Fine-tuning is a learning transfer technique which involves adapting a model that has been pre-trained on a large collection of data to a specific task by continuing training on a target dataset. It allows the reuse of previously learned representations and reduces the need for large quantities of annotated data [19], [29]. There are various variants:

Full tuning: all the parameters of the pre-trained model are updated during training. This allows better specialization, but it is expensive in terms of computation and memory [29].

Freeze tuning: some layers (often the lower ones) are frozen (their weights do not change). Only the top layers (often the classification head or a few top layers) are adjusted. This reduces the computational cost [30].

LoRA tuning: In Low-Rank Adaptation (LoRA) [22], all the pre-trained model weights are frozen and trainable rank decomposition matrices are injected. This approach significantly reduces the number of parameters that need to be trained. LoRA freezes the original weights and applies low-rank correction to the critical linear layers. Let W be the original weight matrix; the updated matrix W′ is defined as,

$$W'=W+\Delta W \text{ (with } \Delta W=AB) \quad (1)$$

If W has dimension d × k, then the dimension of A will be: d × r and the dimension of B will be: r x k. The parameter r is a hyperparameter to be defined (e.g. 4, 8, 16, etc.). When r=1, then A is a column vector and B is a row vector.

II.2 RELATED WORKS

Sentiment analysis in the Algerian dialect is getting more attention in recent years. Researchers at the beginning used classical methods like lexicon-based approaches and machine learning techniques. Currently, deep learning and transformer-based models are becoming popular. In this section, we provide an overview of key works by highlighting the methods and aspects that have been of particular interest to the research community and focusing on works using fine-tuning in the case of AlgD sentiment analysis. The paper of Moudjari et al. [10] addressed the lack of resources in the Algerian dialect by presenting TWIFIL (TWItter proFILing), a collaborative crowdsourcing platform for sentiment and emotion. The sentiment dataset of TWIFIL contains 9437 tweets annotated on 3 classes (negative, positive and neutral).To validate the corpus, several machine learning and deep learning models were evaluated, including SVM, MLP, CNN, LSTM, and multilingual BERT, using different feature representations. Experimental results showed that deep learning models outperformed classical approaches, the CNN model achieved the best performance (76% in terms of accuracy and F1-score).

By [16] focused their sentiment analysis study on only AlgD code-switching comments. They collected more than 36000 YouTube comments, organized into 4 classes (negative, positive, neutral and mixed). The texts are cleaned and normalized. The authors experimented several algorithms: SVM, CNN, LSTM and BiLSTM. In terms of accuracy, the best results were obtained by BiLSTM, it achieved 66.78%, and 60.17% for F1-score by CNN. To remediate the problem of an unbalanced corpus, the authors explored the inclusion of a sentiment lexicon and data augmentation. The insertion of lexicons improved the results for the negative and neutral classes. Abdaoui et al. [21] constructed a corpus of 1.1 million of tweets (15Mo of text) to create DziriBERT, a first pre-trained language model for the AlgD, using the same architecture as BERT base (12 encoders, 12 attention heads and a hidden dimension of 768). The model was trained using the Masked Language Modeling (MLM) task and a WordPiece tokenizer. The overall training time was 10 days. To deploy this pre-trained model, the authors further fine-tuned DziriBERT on an existing AlgD corpus: TWIFIL [10]; as described above. DziriBERT-sentiment achieved 80.3% for accuracy and 79.7% for F1 score.

According to the authors' tests, these results were better than the ones obtained using multilingual models such as mBERT, XLM-R, or Arabic models like AraBERT, QARiB, CamelBERT, and MARBERT. By [25] collected and manually annotated a dataset extracted from YouTube. The corpus contains 45000 texts organized into 3 classes (negative, positive and neutral). The authors initially used a deep learning model LSTM, then experimented FT of various architectures of BERT (base/large) and BERT Arabic (mini/medium/base/large). The results varied from 74.37% to ~81% in terms of accuracy and from 69.76% to ~78% for F1-Score. The results obtained by Bert Arabic (medium/base/large) surpassed those obtained by LSTM or BERT. The best result was obtained by Bert Arabic large (accuracy= 81.74 and F1-score= 78.38). For the best model, parameter's number is 336.7 million with a training time of about 2 hours. The authors tested 11 separate and combined pre-processing operations including classical operations such as stop word deletion and normalization. Some techniques help to increase performance and others reduce it. In contrast, the separation of emojis helped to increase model performance.

By [31] developed a system called AlgBERT, based on new version of Arabic BERT (AraBERT V2- Twitter) for the automatic generation of an annotated corpus. AraBERT Base V2-Twitter has had emoji added to their vocabulary. The AlgBERT model is fine-tuned using a corpus of approximately 54000 texts in Arabic and AlgD (the authors used existing dataset plus data annotated by authors). The classification is of binary polarity (the neutral class is eliminated from dataset). The number of parameters to train is 136 million. The authors use classical text pre-processing, plus the recognition of the polarity of the emoji to improve model performance. After splitting the dataset into train and test, AlgBERT obtained an accuracy of 92.62% and 92.26% of F-measure.The authors [32] collected and normalized two existing datasets of comments in AlgD from Facebook, Instagram, Twitter (now X), and YouTube. The combined corpus contains 12364 comments, annotated into three classes (negative, neutral and positive). The authors applied a complete pre-processing pipeline (with translation of emoji into positive/negative tokens) before training different models: SVM, NB, LR, and Decision Tree, achieving performances between 65% and 74% for accuracy.

They then proceeded to fully fine-tune the pre-trained DziriBERT model. After optimising the hyperparameters, the best result obtained was 82.01% in terms of accuracy using cross-validation, significantly surpassing traditional models. According to. [24] created DZDialect, a large dataset containing 117569 texts from existing datasets augmented by other YouTube comments. The data is organized on 3 classes but the authors used only 2 classes (positive and negative). After text cleaning, they employed various classifiers ML (SVM, KNN, NBM), DL (CNN, LSTM) and transformer models (DistilBERT, AraBERT Base/Mini/Medium, AraGPT-2). In general, transformer-based models gave good results compared with ML and DL except AraGPT-2. AraBERT Base attained an accuracy of 87.96% and 87.95% for F1, in the internal evaluation case. But for the external evaluation, the best results were achieved by AraBERT Mini, with an accuracy of 78.89% and 78.82% for F1 on the Narabizi dataset corpus [33]. To further enhance performance, ensemble learning techniques such as stacking and major voting were employed, specifically with DistilBERT, AraBERT Base, and AraGPT-2.

The results achieved 91% (accuracy and F1 score) using Stacking Ensemble with Logistic Regression on DZDialect. On the Narabizi dataset, using both ensemble methods showed the same performances, with 81.24% accuracy, and 80.57% F1-score. In table 1, we summarise the previous works based on FT, indicating the original model, the number of trained parameters, the volume of data used and their sources (we note, Tw for Twitter, YT: YouTube, Fb: Facebook, In: Instagram), the type of classification; i.e. number of classes; and the best results in term of accuracy (Ac) and F1-score (F1).

Table 1: Summary of works on AlgD sentiment analysis based on fine-tuning.

Model/Ref	model source	time	Ac and F1 score results	# param (million)	classes	Dataset/volume used	Source
DziriBERT/[21]	BERT	10 days	/	110	-	1.1 million tweets	Tw
DziriBERT- sentiment / [21]	mBERT	NA	Ac: 73.6 F1: 72.9	167	3	Twifil corpus (sentiment set)/ 9437	Tw
	XLM-R		Ac: 78.5 F1: 78.0	278			
	AraBERT		Ac: 72.1 F1: 71.3	135			
	QARiB		Ac: 77.7 F1: 77.1	135			
	Camel-BERT-da		Ac: 73.3 F1: 72.7	110			
	Camel-BERT-mix		Ac: 77.1 F1: 76.6	110			
	MARBERT		Ac: 80.1 F1: 79.5	163			
	DziriBERT		**Ac: 80.3 % F1: 79.7 %**	124			
[25]	BERT base	33 min 20 s	Ac: 75.00% F1: 69.79%	~109.5	3	ADArabic: 45000	YT
	BERT large	1 h 50 min	Ac: 74.84% F1: 69.76%	~335.1			
	BERT Arabic mini	2 min 40 s	Ac: 75.27% F1: 70.57%	~11.6			
	BERT Arabic medium	11 min 25 s	Ac: 78.60% F1: 75.21%	~42.1			
	BERT Arabic base	34 min19 s	Ac: 80.02% F1: 76.88%	~110.6			
	BERT Arabic large	1 h 53 min	Ac: 81.74% F1: 78.38%	~336.7			
AlgBERT/[31]	AraBERT V2-Twitter	NA	Ac:92.62 F1: 92.26	136	2	53815	Tw+YT
[32]	DziriBERT	NA	Ac : 82.01% F1 : 81.79%[i]	124	3	12364	Fb+In YT+Tw
[24]	DistilBERT	41 min 14 s	Ac: 85,94% F1: 86,93%	134	2	DZDialect 117569 (used[ii]: 10104)	YT+Tw
	AraBERT Base	1 h18min 15s	Ac: 87,96% F1: 87,95%	110			
	AraBERT Mini	5mn 26s	Ac: 86,03% F1: 86,02%	11			
	AraBERT Medium	24 min 52 s	Ac: 86,7% F1: 86,68%	42			
	AraGPT-2	1 h 25 min 6 s	Ac: 78,90% F1: 78,90%	135			
	Ensemble (LR)	NA	Ac: 91 F1: 91%	-			

Source: Authors, (2026).

After analysing Table 1, it can be seen that the best results were obtained in the case of binary classification. AlgBert [31] reached approximately 92% in both accuracy and F1-score , while a Stacking Ensemble of 3 models fine-tuned by [24] gave around 91%. For a ternary classification, DziriBERT-Sentiment [21] achieved an accuracy of around 80%. Fine-tuning BERT Arabic (Base and Large) [25] gave an accuracy of 81.74%. The best result was obtained by fine-tuning the pre-trained DziriBERT model achieving 82.01% accuracy [32].

III. MATERIALS AND METHODS

The experiments used DziriBERT model and its tokenizer available on Hugging Face[iii] platform. Corpora descriptions will be given below (Table 2). All experiments were conducted on the Google Colab platform, primarily using a T4 GPU, and an A100 GPU or L4 GPU when explicitly mentioned.

The training hyperparameters were fixed at a learning rate of (2×10^{-5}), a batch size of 8, and 5 epochs. For LoRA setting we added the following parameters: $r = 8$, $\alpha = 16$, dropout = 0.1, with targeted application on the Query, Key, and Value modules.

III.1 METHODOLOGY

DziriBERT [21] stands out for its compactness, with approximately 124 million parameters and a vocabulary of 50000 entries, making it one of the smallest models dedicated to AlgD. In this paper, we aim to study the application of DziriBERT to new datasets and to analyse the impact of different fine-tuning strategies on its performance. In this context, two areas of experimentation are possible:

A. DziriBERT[iv], the pre-trained model for the Algerian dialect.
B. DziriBERT-Sentiment[v], the classifier model for AlgD sentiment analysis.

For this study, we chose to work on the DziriBERT-Sentiment model. In the remainder of this paper, we will refer to it simply as DziriBERT. Above all, the dataset must be split into training (Train), validation (Valid) and test (Test) sets. The training and validation sets will be used to train and adjust the model while the test set will be used for model evaluation. We follow the approach illustrated in Figure 1.

1. Firstly, we apply original DziriBERT (without fine-tuning) on the Test portion of the dataset and record the results as baseline performance.
2. Then, we use the training and validation sets (Train and Valid) to fine-tune the DziriBERT model, using three different methods :
   - Full tuning (Full) : all the parameters of the model will be updated during training
   - Freeze tuning (Freeze) : only the parameters of the last layer will be updated during training
   - LoRA tuning (LoRA): all the model weights will be freeze, but we calculate the new matrix ΔW (with ΔW=AB) see section 2.4 for more details.
3. We evaluate the fine-tuned models on the same test set.
4. Next, we extract subsets Sub_Train and Sub_Valid from the training and validation sets, respectively, we then repeat the same process as in step 2 and 3 to fine tune the initial DziriBERT model using these subsets.
5. Finally, we compare the results obtained from the different experiments.

The application of the DziriBERT model explores two experiments: one without pre-processing and one with pre-processing texts from the corpus. The pre-processing step follows the same light pre-processing as described in [21]:

1. replacing all user mentions with @user;
2. replacing all email addresses with mail@email.com;
3. replacing all hyperlinks with https://anonymizedlink.com.

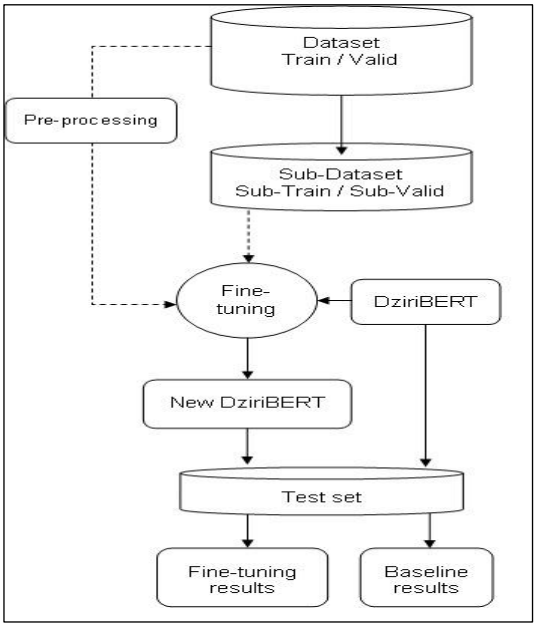

Figure 1: The diagram of the proposed model.
Source: Authors, (2026).

III.2 DATA DESCRIPTION

To evaluate our approach, we use one of the largest public corpora, ADArabic[vi] dataset created by [25], which is dedicated to AlgD SA. To strengthen our study, we also use another corpus collected by [16], referred to here as the Adouane corpus.

III.2.1 ADArabic Dataset

The corpus consists of 45000 annotated comments from YouTube, with five classes of sentiments, from very negative to very positive. We grouped the negative classes (very negative and negative) and the positive classes, to obtain three classes: negative (NEG), neutral (NEU), and positive (POS). First the dataset was split into a training set (80%) and a testing set (20%), and the training set was further divided into Train (80%) and Validation (Valid) (20%). Table 2 presents the basic statistics of resulting dataset.

Table 2: ADArabic corpus description, after regrouping.

Train			**Valid**			**Test**			**Total**
28800			7200			9000			
NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS	45000
10213	6459	12128	2553	1615	3032	3191	2019	3790	

Source: Authors, (2026).

III.2.2 Adouane Dataset

Adouane corpus [16] contains 36120 comments from YouTube. The corpus is organized into three parts: Train, Validation (Valid) and Test. Each part includes four classes (positive, negative, neutral and mix). To maintain consistency with the ternary classification setting, the Mix class was removed, leaving a total of 24,384 comments. A description of the resulting corpus is provided in Table 3.

Table 3: Adouane corpus description, after Mix class removing.

Train			**Valid**			**Test**			**Total**
17304			714			6366			
NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS	24384
4849	5524	6931	167	154	393	1408	1584	3374	

Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

To present the results clearly, they have been organised into sections. For all experiments we measured accuracy (Ac) and macro F1-score (F1) on Test set.

IV.1 BASELINE RESULTS

Table 4 presents the obtained results. Without pre-processing, DziriBERT obtained 76.22% in terms of accuracy and 74.36% for F1-score on the ADArabic corpus, and 73.66% accuracy and 68.34% F1-score on the Adouane corpus. Contrary to our expectations, we have the same finding with or without pre-processing results (see table 4). For that reason, following experiments were performed without pre-processing.

Table 4: DziriBERT result before fine-tuning on 02 datasets.

Dataset	**Without pre-processing**		**With pre-processing**	
	Ac	F1	Ac	F1
ADArabic corpus	76.22%	70.45%	76.22%	70.45%
Adouane corpus	73.66%	68.34%	73.66%	68.34%

Source: Authors, (2026).

IV.2 FINE-TUNING DZIRIBERT ON ADARABIC CORPUS

Regarding the fine-tuning strategies, Full fine-tuning requires training all parameters (+124 million). With freeze tuning, only 2307 parameters need to be trained, which is less than 0.2% of the total DziriBERT parameters. LoRA, on the other hand, trains 444675 parameters, which is equivalent to 0.36% compared to full fine-tuning parameters number. This is valid for all the following experiments.

IV.2.1 DziriBERT Fine-Tuning Results on ADArabic (Full Corpus)

Table 5 shows the DziriBERT FT results, on the whole ADArabic corpus. The experiments were running using L4 GPU. We can see that all three FT techniques (Full, Freeze and LoRA) improved the baseline results in terms of both accuracy and F1 scores. Full FT achieved 82.57% accuracy and 79.48% F1-score after 1 hour (hr) and 42 minutes (mn) of training, while Freeze achieved 79.51% accuracy and 75.54% F1-score in 1 hour and 35 minutes. LoRA achieved the best performance with 83.26% accuracy and 80.94% F1-score in the shortest training time of 1 hour and 24 minutes. When analysing the confusion matrix (see Figure 2), it is obvious that it is always the neutral class that is wrongly predicted, this is probably due to the imbalanced data.

Table 5: FT results on ADArabic corpus using GPU L4.

	# params	**Time of training**	**Ac**	**F1**
Full	124 443 651	1hr 42mn	82.57%	79.48%
Freeze	2 307	1hr 35mn	79.51%	75.54%
LoRA	444 675	1hr 24mn	83.26%	80.94%

Source: Authors, (2026).

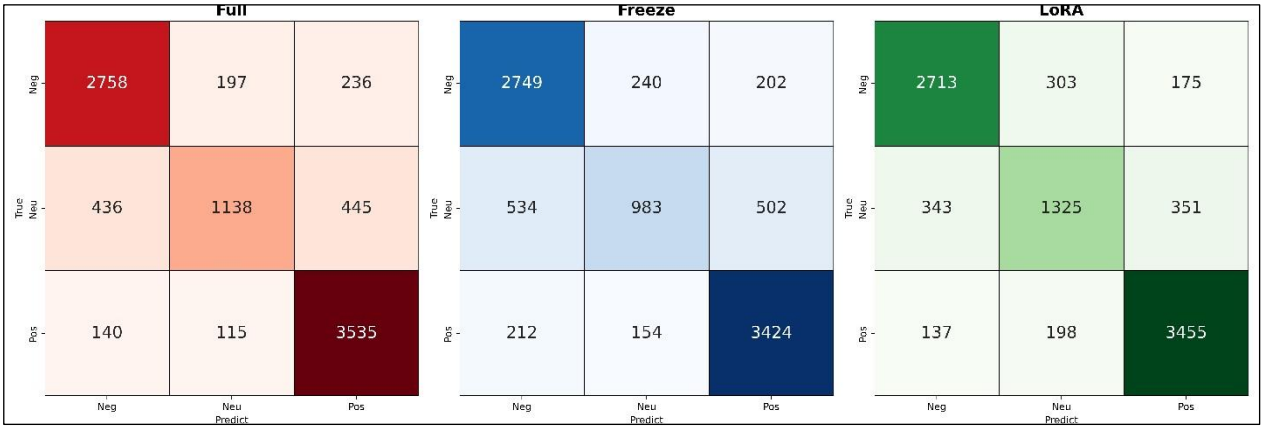

Figure 2: Confusion matrix of ADArabic fine-tuning models.
Source: Authors, (2026).

IV.2.2 Impact of Dataset Size on Fine-Tuning Performance

To explore the final objective of this paper, we examined the results obtained by fine-tuning DziriBERT using subsets of the training (Train) and validation (Valid) sets. For each experiment, we tested several values of $N1$ and $N2$ representing the sizes of Train and Valid subsets respectively. The choice of N1 was almost arbitrary while N2 was generally set to 20% of N1. We started with N1=1000 and a step size of 500. If the variations in the results did not exceed 1 point, we moved on to larger step size such as: 2500 or 5000.

Table 6 summarises the results of all the experiments on the Test set. The last row represents the results obtained using the complete dataset. The best values for each technique are highlighted in bold, and the best values for each subset are underlined. The results of all experiments outperform the baseline in terms of both accuracy and F1-Score. In general, the best results are achieved using the full dataset.

Table 6: Impact of dataset size on performance of FT on ADArabic corpus.

		Full		**Freeze**		**LoRA**	
N1	N2	Ac	F1	Ac	F1	Ac	F1
1000	200	80.39%	76.82%	76.66%	71.54%	80.62%	77.49%
1500	300	80.41%	77.79%	76.81%	71.91%	79.20%	75.58%
2000	400	81.23%	78.34%	76.86%	72.06%	81.23%	78.40%
2500	500	80.87%	77.38%	76.94%	72.30%	80.89%	77.91%
3000	600	80.86%	77.43%	77.04%	72.43%	81.42%	78.67%
5000	1000	80.07%	78.27%	77.47%	73.13%	78.96%	75.27%
10000	2000	81.29%	78.37%	78.16%	73.91%	80.99%	77.92%
15000	3000	82.01%	79.29%	78.66%	74.42%	82.31%	80.02%
20000	4000	82.34%	78.92%	79.03%	74.92%	82.72%	80.30%
25000	5000	**82.63%**	**80.16%**	79.32%	75.36%	82.82%	80.36%
28800	7200	82.57%	79.49%	**79.51%**	**75.54%**	**83.26%**	**80.94%**

Source: Authors, (2026).

For Full fine-tuning (Full), when N1 = 1000, accuracy reached 80.39% and F1-score 76.82%, corresponding to improvements of 4.17 and 6.37 points respectively, compared to the baseline. Full reached its maximum at N1 = 25000, where accuracy reached 82.63% (an improvement of 6.41 points) and F1 reached 80.16% (an improvement of 9.71 points). With Freeze tuning, accuracy increased from 76.66% to 79.51% and F1 from 71.54% to 75.54%, showing maximum improvement of 3.29 points for accuracy and 5.09 points for the F1-score when using the complete dataset. Using LoRA at N1 = 1000, accuracy reached 80.62% and F1 77.49%, corresponding to improvements of 4.40 and 7.04 points respectively. LoRA reached its maximum performance with the whole dataset, achieving 83.26% accuracy (an improvement of 7.04 points) and 80.94% F1 (an improvement of 10.49 points). The results for Freeze are the lowest, its best results using the whole dataset are still lower than those obtained by Full or LoRA using only 1200 texts (N1=1000 and N2=200). The latter achieved around 80% for accuracy and 77% for F1. Using a subset of 2400 texts, Full and LoRA techniques achieved a local maximum of Accuracy ≈81% and F1≈78, which means an improvement of baseline results by 5 points in terms of accuracy and around 8 points for F1. In most cases, LoRA's performance are better than Full ones (see Table 6 and Figure 3).

IV.3 DZIRIBERT FINE-TUNING ON ADOUANE CORPUS

On the Adouane corpus, we explored the same experiments as in the previous section, and we summarized all the results in Table 7. The last row corresponds to the whole-dataset results. The number of parameters for each experiment does not change. For this run, we started with N1 = 1000 with a step size of 1000. After N1 = 4000, we jump directly to the whole dataset because the validation set contains only 714 texts. This is due to the removal of the Mix class (see the Data description section). We have almost the same findings as in the previous section. The three FT techniques improved the baseline results (see Table 7). Using Full technique, accuracy started from 81.95% to 84.70%, F1 from 78.98% to 81.56%. Therefore, compared to baseline, the improvements can reach 11.04 and 13.22 points for accuracy and F1, respectively, using the whole dataset, in the training time of 20 minutes using an A100 GPU.

Regarding Freeze, accuracy started from 75.12% to 79.89%, F1 from 70.71% to 76.61%. Therefore, the improvements can reach 6.23 and 8.27 points. Processing the whole dataset took only 8 minutes using an A100 GPU. LoRA strategy achieved the best results, an accuracy of 85.28% and 82.52% for F1, in the training time of 15 minutes and 59.49 seconds using an A100 GPU. The enhancements started from 8.53 points to 11.62 for accuracy and F1 from 10.76 points to 14.18 points. In this corpus, we found that using subset of 1200 texts (N1=1000 and N2=200) gives results close to the original results reported by [21] the creator of DziriBERT when using Full or LoRA fine-tuning techniques. On the other hand, using a subset of N1 equal to 2000 or 3000, Full or LoRA fine-tuning can achieve results close to the best performance obtained so far by [32] based on the fine-tuning of the pre-trained DziriBERT model (see Table1).

Table 7: Impact of dataset size on FT performance on the Adouane corpus.

Technique		**Full**		**Freeze**		**LoRA**	
# params		124 443 651		2307		444 675	
N1	N2	Ac	F1	Ac	F1	Ac	F1
1000	200	81.95%	78.98%	75.12%	70.71%	82.19%	79.10%
2000	400	82.42%	78.84%	75.79%	71.67%	82.50%	79.31%
3000	600	82.08%	79.21%	76.81%	73.03%	83.95%	80.96%
4000	714	83.62%	80.28%	77.18%	73.45%	83.98%	81.04%
17304	714	**84.70%**	**81.56%**	**79.89%**	**76.61%**	**85.28%**	**82.52%**

Source: Authors, (2026).

IV.4 DISCUSSION

DziriBERT model gave moderate results (baseline), around 75% on ADArabic corpus and a little less on Adouane Corpus. For both datasets, looking at the accuracy and F1-score graphs (see figure 4, where O.W is used to designate Original Work and B.L baseline), Freeze FT improved baseline results in an increasing and monotonic manner in relation to N1. Full FT and LoRA FT had results close to each other. They alternated between growth and decline. They increased baseline by 6-14 points (around 10 points) when using the whole dataset. Sometimes LoRA slightly outperformed Full FT, while requiring significantly fewer parameters and time.

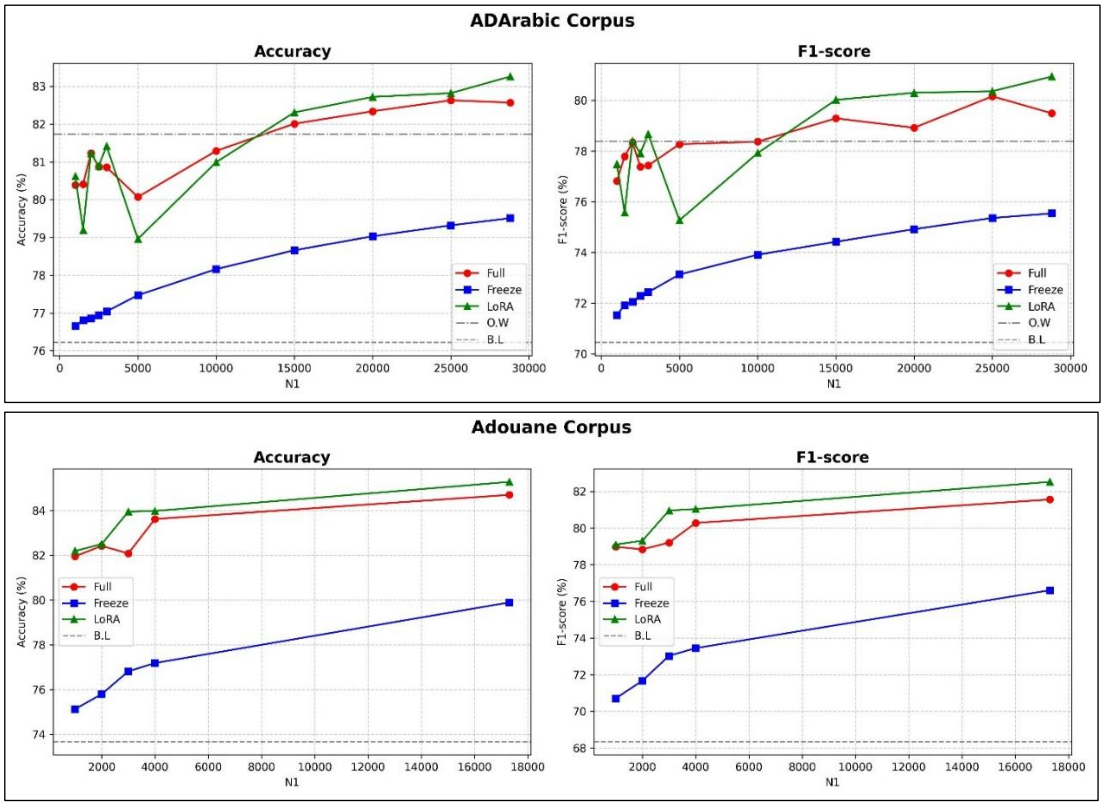

Figure 3: Accuracy and F1-score variation according to N1 using Full, Freeze and LoRA fine-tuning. (O.W:original work and B.L: baseline).

Source: Authors, (2026).

Certainly, with Freeze FT, we had a very small number of parameters to train, and although it improve the baseline results, but they are still very limited and inferior to those given by Full or LoRA FT. Here, Freeze retrained only the last layer (the layer responsible for classification). Even if we use the whole dataset, the Freeze results remain lower than those obtained by retraining all parameters with subset of 1200 texts, which means that the word embedding weights of AlgD in DziriBERT must be readjusted, maybe by other words from different social networks. Noting that DziriBert was trained on one million Twitter texts. The latter is not the most widely used network by Algerians (see Figure 4) but it ranks last [34].

For a ternary classification using the DziriBERT model: [21] achieved an accuracy of 80.3% using a dataset of 9437 Twitter texts, [32] achieved 82.01% using FT of DziriBERT with 12364 texts from Twitter, YouTube, Facebook and Instagram. Our work achieved 83.26% on ADArabic corpus that contains 45000 texts from YouTube, thus exceeding the results of the authors of the original work, who based their research on FT several BERT models.-On ADArabic corpus, the original work by [25] achieved 81.74% accuracy and 78.38% F1-score, based on fine-tuning six BERT models variants with multiple pre-processing treatments. On the same corpus, our approach reached 83.26% accuracy and 80.94% F1-score using only DziriBERT fine-tuning without any pre-processing.

We achieved also 85.28% accuracy on Adouane corpus (24384 texts from YouTube), this result represents the best performance obtained for AlgD SA in ternary classification. This finding indicates that DziriBERT is well adapted to Algerian sentiment analysis and probably other NLP tasks. On both datasets, we found that Full FT and LoRA FT significantly improved the models, by using the complete dataset or sub-dataset. But LoRA combines the advantages of both Full and Freeze techniques, providing significant performance improvements like Full while requiring a very small number of parameters to be retrained similar to Freeze. In addition, LoRA results usually outperforms those of Full FT, so we can adopt LoRA as the best FT strategy. On the other hand, LoRA also reduces energy consumption. This helps to preserve the environment.

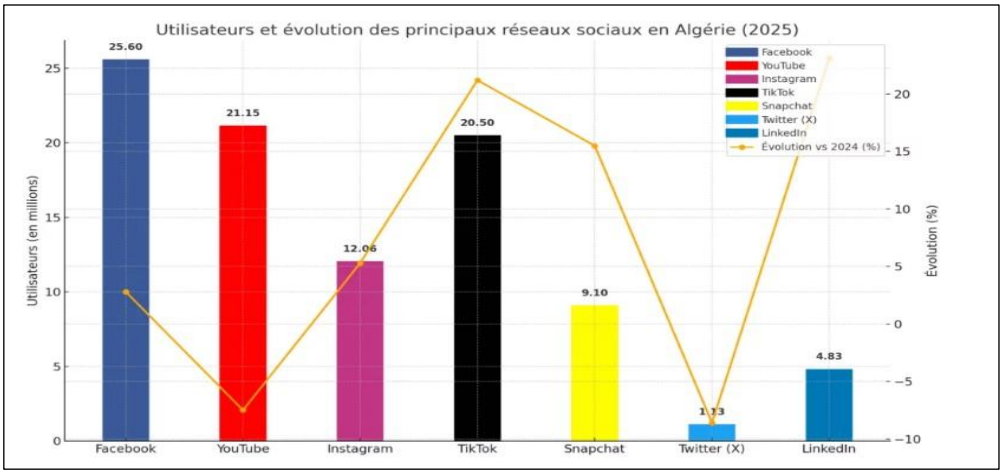

Figure 4: Users and the evolution of social networks in Algeria.
Source: [34].

The experiments conducted in this research suggest that using a small dataset for FT can produce satisfactory results. This emphasises that data quality can sometimes take precedence over quantity. Small data means fewer instructions need to be run in less time, which also helps to preserve the environment. As future work, we plan to: (i) Test DziriBZRT on other corpora. (ii) Fine tune the pre-trained DziriBERT model with texts from others sources such as Facebook and Instagram (in addition to Twitter). (iii) Deal with others Sentiment Analysis challenges such as: sarcasm or negation treatment.

V. CONCLUSIONS

In this paper, we have explored the sentiment analysis of Algerian dialect using three techniques of fine-tuning (Full, Freeze and LoRA tuning) on DziriBERT. The study is conducted on two existing datasets; ADArabic and Adouane corpus. The results of all experiments are better than the baseline, i.e. when using DziriBERT without FT. On both datasets, we found that Full and LoRA significantly improved the models, by using the complete dataset or sub-dataset. But LoRA combines the advantages of both Full and Freeze techniques, i.e. significant performance improvements like Full and a very small number of parameters to be retrained like Freeze. In addition, LoRA results usually outperforms Full ones, so we can adopt LoRA as the best FT strategy. On the other hand, LoRA has an impact on energy consumption. This helps to preserve the environment.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Salima Brachemi-Meftah.
Methodology: Salima Brachemi-Meftah and Fatiha Barigou.
Investigation: Salima Brachemi-Meftah.
Discussion of results: Salima Brachemi-Meftah,
Writing – Original Draft: Salima Brachemi-Meftah.
Writing – Review and Editing: Salima Brachemi-Meftah and Fatiha Barigou.
Resources: Salima Brachemi-Meftah and Fatiha Barigou.
Supervision: Fatiha Barigou.
Approval of the final text: Salima Brachemi-Meftah and Fatiha Barigou.

VII. ACKNOWLEDGMENTS

We would like to thank Jean-Philippe Bernardy, Wafia Adouane, and Samia Touileb, for providing us their dataset, which enabled us to conduct this research.

VIII. REFERENCES

[1] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, 'A comprehensive survey of arabic sentiment analysis', Information Processing & Management, vol. 56, no. 2, pp. 320–342, Mar. 2019, doi: 10.1016/j.ipm.2018.07.006.

[2] A. Ghallab, A. Mohsen, and Y. Ali, 'Arabic Sentiment Analysis: A Systematic Literature Review', Applied Computational Intelligence and Soft Computing, vol. 2020, pp. 1–21, Jan. 2020, doi: 10.1155/2020/7403128.

[3] B. Liu, 'Sentiment Analysis and Opinion Mining', Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[4] N. O. Adelakun and A. A. Baale, 'Sentiment analysis of financial news using the BERT model', ITEGAM-JETIA, vol. 10, no. 48, pp. 21–27, Jul. 2024, doi: 10.5935/jetia.v10i48.1029.

[5] I. Villanueva-Miranda, Y. Xie, and G. Xiao, 'Sentiment analysis in public health: a systematic review of the current state, challenges, and future directions', Frontiers in Public Health, vol. 13, p. 1609749, Jun. 2025, doi: 10.3389/fpubh.2025.1609749.

[6] A. M. Alayba, 'Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends', Computers, vol. 14, no. 11, Nov. 2025, doi: 10.3390/computers14110497.

[7] H. A. Alkaabi, A. kadhim Jasim, and A. Darroudi, 'Arabic NLP: A Survey of Pre-Processing and Representation Techniques', Journal of Computer Science, Information Technology and Telecommunication Engineering, vol. 6, no. 2, pp. 876–890, Sep. 2025, doi: 10.30596/jcositte.v6i2.25562.

[8] S. Brachemi-Meftah and F. Barigou, 'Algerian Dialect Sentiment Analysis: Sate of Art', in 2020 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt: IEEE, Nov. 2020, pp. 1–7. doi: 10.1109/ACIT50332.2020.9300060.

[9] A. Dahou et al., 'A Survey on Dialect Arabic Processing and Analysis: Recent Advances and Future Trends', ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 24, no. 8, p. 84:1-84:45, Aug. 2025, doi: 10.1145/3747290.

[10] L. Moudjari, K. Akli-Astouati, and F. Benamara, 'An Algerian Corpus and an Annotation Platform for Opinion and Emotion Analysis', presented at the Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, May 2020, pp. 1202–1210. Accessed: Jun. 05, 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.151

[11] M. Al-Ayyoub, S. B. Essa, and I. Alsmadi, 'Lexicon-based sentiment analysis of Arabic tweets', International Journal of Social Network Mining, vol. 2, no. 2, pp. 101–114, Jan. 2015, doi: 10.1504/IJSNM.2015.072280.

[12] K. Arun and A. Srinagesh, 'Multilingual twitter sentiment analysis using machine learning', International Journal of Electrical and Computer Engineering (IJECE), vol. 10, no. 6, Art. no. 6, Dec. 2020, doi: 10.11591/ijece.v10i6.pp%p.

[13] P. Lin and X. Luo, 'A Survey of Sentiment Analysis Based on Machine Learning', in Natural Language Processing and Chinese Computing, X. Zhu, M. Zhang, Y. Hong, and R. He, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 372–387. doi: 10.1007/978-3-030-60450-9_30.

[14] Z. Nassr, N. Sael, and F. Benabbou, 'Machine Learning for Sentiment Analysis: A Survey', in Innovations in Smart Cities Applications Edition 3, M. Ben Ahmed, A. A. Boudhir, D. Santos, M. El Aroussi, and İ. R. Karas, Eds., in Lecture Notes in Intelligent Transportation and Infrastructure. Cham: Springer International Publishing, 2020, pp. 63–72. doi: 10.1007/978-3-030-37629-1_6.

[15] N. Omar, M. Albared, T. Al-Moslmi, and A. Al-Shabi, 'A Comparative Study of Feature Selection and Machine Learning Algorithms for Arabic Sentiment Classification', in Information Retrieval Technology, A. Jaafar, N. Mohamad Ali, S. A. Mohd Noah, A. F. Smeaton, P. Bruza, Z. A. Bakar, N. Jamil, and T. M. T. Sembok, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 429–443. doi: 10.1007/978-3-319-12844-3_37.

[16] W. Adouane, S. Touileb, and J.-P. Bernardy, 'Identifying Sentiments in Algerian Code-switched User-generated Comments', presented at the Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, May 2020, pp. 2698–2705. Accessed: Jun. 05, 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.328

[17] M. H. Shakeel and A. Karim, 'Adapting deep learning for sentiment classification of code-switched informal short text', in Proceedings of the 35th Annual ACM Symposium on Applied Computing, in SAC '20. Brno, Czech Republic: Association for Computing Machinery, Mar. 2020, pp. 903–906. doi: 10.1145/3341105.3374091.

[18] A. C. Mazari and A. Djeffal, 'Deep Learning-Based Sentiment Analysis of Algerian Dialect during Hirak 2019', in 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), Feb. 2021, pp. 233–236. doi: 10.1109/IHSH51661.2021.9378753.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[20] C. Sun, X. Qiu, Y. Xu, and X. Huang, 'How to Fine-Tune BERT for Text Classification?', 2019, arXiv. doi: 10.48550/ARXIV.1905.05583.

[21] A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, 'DziriBERT: a Pre-trained Language Model for the Algerian Dialect', arXiv:2109.12346 [cs], Sep. 2021, Accessed: Jan. 17, 2022. [Online]. Available: http://arxiv.org/abs/2109.12346

[22] E. J. Hu et al., 'LoRA: Low-Rank Adaptation of Large Language Models', Oct. 16, 2021, arXiv: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.

[23] L. Martin et al., 'CamemBERT: a Tasty French Language Model', in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 7203–7219. doi: 10.18653/v1/2020.acl-main.645.

[24] D. Boughareb, R. Boughareb, S. Hallaci, M. R. E. Boukherouba, and H. Seridi, 'Addressing the Complexity of Dialectal Arabic: An Enhanced Encoder-Decoder Ensemble Approach for Optimized Sentiment Analysis', ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 24, no. 6, pp. 1–21, Jun. 2025, doi: 10.1145/3735972.

[25] Z. Benmounah, A. Boulesnane, A. Fadheli, and M. Khial, 'Sentiment Analysis on Algerian Dialect with Transformers', Applied Sciences, vol. 13, no. 20, p. 11157, Jan. 2023, doi: 10.3390/app132011157.

[26] H. Saâdane, 'Le traitement automatique de l'arabe dialectalisé: aspects méthodologiques et algorithmiques."Automatic processing of dialectical Arabic: methodological and algorithmic aspects"', PhD Thesis, Université Grenoble Alpes, 2015.

[27] W. X. Zhao et al., ‘A Survey of Large Language Models’, Mar. 11, 2025, arXiv: arXiv:2303.18223. doi: 10.48550/arXiv.2303.18223.

[28] W. Antoun, F. Baly, and H. Hajj, ‘AraBERT: Transformer-based Model for Arabic Language Understanding’, in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, Eds., Marseille, France: European Language Resource Association, May 2020, pp. 9–15. Accessed: Aug. 04, 2025. [Online]. Available: https://aclanthology.org/2020.osact-1.2/

[29] J. Howard and S. Ruder, ‘Universal Language Model Fine-tuning for Text Classification’, May 23, 2018, arXiv: arXiv:1801.06146. doi: 10.48550/arXiv.1801.06146.

[30] M. E. Peters, S. Ruder, and N. A. Smith, ‘To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks’, Jun. 11, 2019, arXiv: arXiv:1903.05987. doi: 10.48550/arXiv.1903.05987.

[31] K. Hamadouche, K. Z. Bousmaha, M. A. Bekkoucha, and L. Hadrich-Belguith, ‘AlgBERT: Automatic Construction of Annotated Corpus for Sentiment Analysis in Algerian Dialect’, ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 22, no. 12, pp. 1–17, Dec. 2023, doi: 10.1145/3632948.

[32] F. Bougamouza and S. Hazmoune, ‘Sentiment analysis of customer reviews for Algerian dialect using the DziriBERT model’, International Journal of Data Analysis Techniques and Strategies, Aug. 2024, Accessed: Jul. 21, 2025. [Online]. Available: https://www.inderscienceonline.com/doi/10.1504/IJDATS.2024.140647

[33] S. Touileb and J. Barnes, ‘The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus’, in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 3700–3712. doi: 10.18653/v1/2021.findings-acl.324.

[34] M. Amokrane, ‘Réseaux sociaux en Algérie : tendances et chiffres Clés de 2025’, Maghreb Émergent. Accessed: Dec. 25, 2025. [Online]. Available: https://maghrebemergent.news/fr/reseaux-sociaux-en-algerie-tendances-et-chiffres-cles-de-2025/

Note:

---

i Calculated using confusion matrix.
ii Calculated by author of this paper.
iii https://huggingface.co/ (Accessed April 2025)
iv https://huggingface.co/alger-ia/dziribert (Accessed April 2025)
v https://huggingface.co/alger-ia/dziribert_sentiment (Accessed April 2025)
[vi] Available at : https://data.mendeley.com/datasets/zzwg3nnhsz/1 Accessed : April, 2025