Unveiling the Effectiveness of Contextual Embeddings in Sentiment Analysis on Moroccan Darija: A Comparative Study with Traditional Techniques

Meriem Amnay, Abdelfattah Toulaoui, Mourad Jabrane and Imad Hafidi

***Abstract*—Sentiment analysis for dialectal languages like the Moroccan Darija present major challenges due to linguistic variability and the lack of annotated resources. This paper compares traditional embedding techniques (FastText, Word2Vec) with a contextual transformer model (DarijaBERT) specifically trained for Moroccan Darija. We evaluate these embeddings using multiple machine learning classifiers, including Logistic Regression, SVM, Random Forest, and MLP. Our results demonstrate that DarijaBERT consistently outperforms traditional embeddings, particularly when combined with neural classifiers such as MLP. These findings confirm the effectiveness of contextual embeddings for sentiment analysis in under-resourced dialects such as Darija, offering a foundation for future NLP research in similar languages.**

***Index Terms*—Sentiment Analysis, Transformers, BERT, DarijaBERT, AraBERT, Moroccan Darija, Arabic Dialects, NLP.**

I. Introduction

SENTIMENT analysis in Natural Language Processing (NLP) is a crucial technique aimed at extracting and quantifying emotions from text, enabling the classification of sentiments as positive, negative, or neutral. The process is as follows: first we can use feature extraction techniques such as Word2vec and TF-IDF, which convert words into high-dimensional vectors to capture semantic meaning [1], then we employ either Machine learning models, for example, Support Vector Machines (SVM) or deep neural networks to classify the sentiments effectively. Some approaches even addressed challenges like the lack of labeled data [2]. The application of sentiment analysis spans multiple domains, including social media platforms like Twitter, where it helps gauge public opinion and customer experiences [3] . By utilizing advanced NLP techniques, researchers can analyze vast amounts of unstructured data, leading to improved accuracy and insights in understanding human emotions [4].

The launching of transformer-based models marked a significant advancement in sentiment analysis and NLP more broadly. Transformers, introduced by Vaswani [65], brought forth a novel mechanism known as self attention, which allows models to weigh the importance of each word in a sentence based on its relationship to other words. This mechanism enables transformers to capture both local (word-level) and global (sentence- or document-level) context, offering a more comprehensive understanding of a text's meaning. By focusing on the most relevant parts of a sentence, transformers can discern sentiment more accurately, leading to substantial performance improvements in various NLP tasks, including sentiment analysis.

In this paper, we conduct a comparative study of transformer models for sentiment analysis in Moroccan Darija, a dialect of Arabic that presents unique challenges due to its primarily spoken nature, lack of standardization, and significant regional variations. Unlike Modern Standard Arabic (MSA), which is used in formal writing and has a relatively uniform structure, Darija varies across regions in terms of vocabulary, grammar, and influences from other languages such as Berber, French, and Spanish. These characteristics of Darija introduce complexities in NLP tasks that are not encountered in MSA.

To address these challenges, we evaluate six transformer models: BERT Multilingual, DarijaBERT, DarijaBERT Arabizi, MARBERT, AraBERT, and CAMEL BERT, assessing their effectiveness in performing sentiment analysis on a custom Moroccan Darija dataset. This evaluation explores how well each model can handle the intricacies of Darija, including its informal, unstandardized nature and regional diversity, by analyzing their performance on sentiment classification tasks. Our goal is to determine which model provides the most accurate and reliable sentiment predictions in this dialect and to highlight the benefits of using dialect-specific models over more general, multilingual models.

The structure of the paper is organized as follows: In Section II, we discuss related works, focusing on recent advancements in sentiment analysis using transformer models and the specific challenges associated with dialectal Arabic, particularly Moroccan Darija. Section III provides an in-depth look at the experimental evaluation, where we detail the configuration of our experiments, including the tools, libraries, and datasets used. This section also

Manuscript received October 7, 2024; revised May 23, 2025.

The authors contributed equally to this work.

Meriem Amnay is a Ph.D candidate at LIPIM laboratory, University Sultan Moulay Slimane, Beni-Mellal 23000, Morocco (Phone 00212 6 62 17 40 37 e-mail: amnay.meriem@gmail.com).

Abdelfattah Toulaoui is a Ph.D candidate at LIPIM laboratory, University Sultan Moulay Slimane, Beni-Mellal 23000, Morocco (e-mail:abdelfattah.toulaoui1@usms.ma).

Mourad Jabrane is a Ph.D candidate at LIPIM laboratory, University Sultan Moulay Slimane, Beni-Mellal 23000, Morocco (e-mail: mourad.jabrane@usms.ac.ma).

Imad Hafidi is a professor and researcher at LIPIM laboratory, University Sultan Moulay Slimane, Beni-Mellal 23000, Morocco (e-mail: i.hafidi@usms.ma).

explains the preprocessing and tokenization steps necessary for adapting transformer models to the Darija dialect. Section IV covers the comparative study, including our methodology and results. Here, we evaluate various transformer models, describe the evaluation metrics used, and present a detailed analysis of model performance. This section also addresses the unique challenges encountered in analyzing sentiment in Darija and proposes directions for future research to improve the effectiveness of these models. Finally, Section Vconcludes the paper by summarizing the main findings of our comparative study. We highlight the contributions of this research in advancing sentiment analysis for low-resource dialects and suggest potential areas for further exploration, particularly in enhancing model adaptability and accuracy for dialect-specific tasks.

II. Preliminearies

To understand the application of sentiment analysis on Moroccan Darija, it's essential to explore foundational concepts in natural language processing. This section provides an overview of transformer models, the basis for advanced language models, and discusses how these models are adapted for specific tasks. Additionally, we examine the unique linguistic characteristics of Moroccan Darija, a dialect that poses distinct challenges in NLP due to its variations and lack of standardization.

A. Transformers

Transformer models have transformed the field of natural language processing (NLP) by offering a powerful and flexible approach to understanding text. Introduced by Vaswani et al. in 2017, transformers leverage a self-attention mechanism, allowing them to weigh the importance of different words in a sequence based on their contextual relationships. This enables transformers to process entire sequences in parallel, making them more efficient and effective than previous sequential models like recurrent neural networks (RNNs). By capturing both local and global context, transformers have set new benchmarks in various NLP tasks, from machine translation to sentiment analysis. Their versatility and ability to model complex dependencies in language make transformers the backbone of many advanced language processing systems today.

B. Language Model

In the realm of NLP, language models are specialized versions of transformer models that have been fine-tuned for specific language tasks. These models are pre-trained on large corpora to learn general language patterns and then adapted to perform specific tasks like text generation, translation, and sentiment analysis. Fine-tuning allows a general transformer model to develop specialized skills by training it on task-specific data. As a result, language models are essentially transformers that have been optimized for particular applications, enabling them to handle nuances and subtleties within specific domains or languages. This adaptability is key to their effectiveness, especially when applied to dialects or specialized linguistic contexts like Moroccan Darija.

C. Dialectal Variation in Moroccan Darija

Moroccan Darija exhibits significant linguistic variability, even within different regions of Morocco, due to the influence of several languages, including Berber, French, and Spanish. These external influences have contributed to Darija's unique vocabulary, grammatical structures, and phonetics, making it distinct from both Modern Standard Arabic (MSA) and other Arabic dialects. One of the most striking features of Darija is the absence of a standardized writing system. Unlike MSA, which follows formal grammatical rules and standardized spelling, Darija is primarily an oral dialect. As a result, when Darija is written, people often rely on phonetic transcriptions that reflect spoken language rather than adhering to any fixed orthographic conventions.

Furthermore, individuals frequently write Darija using both the Arabic script and Latin characters, the latter being referred to as Arabizi.This dual usage of writing systems adds considerable complexity to the process of data collection, as texts written in Darija are highly inconsistent in their format, spelling, and grammar. The lack of standardization makes it challenging to compile cohesive datasets for tasks like sentiment analysis, where linguistic uniformity is essential for training models effectively.

In this study, we address the complexities of both Darija in Arabic script and Darija in Arabizi by incorporating models specifically designed to handle these variations, such as DarijaBERT Arabizi, which is trained on text written in Latin characters. The presence of multiple writing styles within the same language introduces additional challenges for both tokenization and text normalization. Tokenization, the process of breaking down text into smaller units (such as words or subwords) for model processing, becomes difficult when faced with non-standard spellings and mixed scripts. Similarly, normalization—ensuring consistency in text by converting different forms of the same word or phrase into a uniform representation—requires specialized approaches to manage the variability inherent in Darija.

These challenges highlight the importance of developing specialized models that can accommodate the unique characteristics of Moroccan Darija, particularly when it is written in different scripts. Models like DarijaBERT Arabizi, which are fine-tuned on datasets containing Arabizi text, play a crucial role in effectively handling sentiment analysis tasks in this dialect. By taking into account both the Arabic script and Latin-based transcriptions, these models are better equipped to manage the diversity of written forms in Darija, thereby improving performance in NLP tasks that rely on accurate understanding and processing of the dialect.

III. Related works

This section reviews previous studies relevant to sentiment analysis, particularly in the context of Arabic and its dialects. We begin by exploring how language models have been applied to sentiment analysis, followed by a look into the specific challenges posed by Arabic and its diverse dialectal variations. These insights provide a foundation for understanding the complexities involved in adapting sentiment analysis models to dialectal Arabic. Previous research in Arabic sentiment analysis has primarily focused on Modern Standard Arabic (MSA), leaving dialectal varieties underexplored. Early studies employed machine learning algorithms with basic lexical features. More recent work integrates word embeddings and deep learning methods to improve performance. However, few studies have investigated the performance of contextual embeddings for Moroccan Darija specifically, highlighting a significant gap that this study aims to address.

A. Language models for Sentiment Analysis

Recent advancements in machine learning and deep learning have substantially enhanced the field of natural language processing (NLP) across a variety of applications, including machine translation [65], [61], entity resolution [5], [6], [8], [7], sentiment analysis [62], [67], [55], [56], question answering systems [63], [64], and Arabic news classification and detection [10], [11].

Transformers models have revolutionized natural language processing (NLP) by allowing efficient parallel processing of sequential data. Recent improvements, such as deeper transformer architectures, have further boosted performance by using advanced layer connections and normalization techniques [60]. The Transformer remains a foundation of modern deep learning, driving significant progress across multiple domains[12]. Transformers have emerged as a powerful architecture for sentiment analysis, leveraging their ability to capture contextual relationships in text. The BMT-Net model combines feature-based and fine-tuning approaches, achieving superior performance in sentiment analysis tasks by learning universal representations across multiple tasks [18]. Additionally, the Modulated Fusion model integrates linguistic and acoustic inputs, demonstrating effectiveness across various datasets, thus enhancing emotion recognition capabilities[68]. For more complex scenarios, such as aspect-based sentiment analysis, the Transformer-based Multi-aspect Modeling scheme effectively identifies sentiments associated with multiple aspects within a single sentence, outperforming traditional models [20]. Furthermore, multimodal sentiment analysis benefits from Transformer architectures, as seen in the MEDT and Gate-Fusion Transformer models, which address long-term dependencies and inter-modal interactions, leading to improved emotional understanding from diverse data sources [22], [23]. Collectively, these advancements highlight the versatility and robustness of Transformers in sentiment analysis

Fine-tuning is a crucial process in natural language processing (NLP) that allows pre-trained models to be adapted to specific tasks or domains by training them on a smaller, labeled dataset. In the context of sentiment analysis, transformer models that have been pre-trained on large general corpora are further fine-tuned on sentiment-specific datasets. These datasets typically contain text labeled with sentiment categories, such as positive, negative, or neutral, which helps the model to learn how to accurately classify the emotional tone of new, unseen text. Additionally, fine-tuning BERT for sentiment analysis has proven effective, with experiments indicating that BERT outperforms traditional models like GloVe and FastText on Vietnamese review datasets [13].

Language models play a crucial role in sentiment analysis (SA), leveraging both traditional and advanced methodologies to interpret emotions in text. Recent advancements in large language models (LLMs) have shifted the focus towards emotionally-agnostic sentiment analysis, which emphasizes textual cues over physiological indicators, enhancing the accuracy of emotional response generation [17]. applications.

Overall, these developments underscore the evolving landscape of sentiment analysis through innovative language model applications [21].

B. Sentiment Analysis in Arabic

Sentiment analysis can be effectively performed using both machine learning and deep learning techniques, each with its own strengths and weaknesses. Traditional machine learning models[57], [58] like Naive Bayes, SVM[59], and Decision Trees have shown good performance, particularly with smaller datasets, but they require extensive feature extraction, which can be time-consuming [32]. In contrast, deep learning models, such as CNNs, RNNs, and LSTMs, automate feature extraction and excel in handling large volumes of data, often yielding superior accuracy and recall [28], [29]. Recent studies indicate that architectures like BERT and LSTM outperform traditional methods, although they may require longer training times [30]. Furthermore, combining multiple algorithms can enhance accuracy, making it a viable strategy for diverse text types [31]. Ultimately, the choice between machine learning and deep learning for sentiment analysis depends on the specific application, dataset size, and required accuracy[32].

Sentiment analysis in Arabic and Moroccan Darija can effectively utilize both machine learning and deep learning techniques. Research indicates that deep learning models, such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), outperform traditional shallow learning classifiers like Random Forest and Decision Trees in various sentiment analysis tasks, particularly when dealing with large datasets [24], [26]. In the context of Moroccan dialect, the use of AraBERT for word embedding has shown promising results, with BiLSTM achieving high accuracy in both binary and multi-class classifications [27]. Additionally, studies focusing on Arabic sentiment analysis have highlighted the importance of preprocessing and feature extraction tailored to dialectal variations, which enhances model performance [25], [27]. Overall, the combination of deep learning architectures and tailored preprocessing strategies is crucial for effectively analyzing sentiments in Arabic and Moroccan Darija.

C. Word Embedding representations for sentiment analysis

Word embedding techniques have become a base in sentiment analysis, offering a means to transform textual data into numerical vectors that can be processed by machine learning and deep learning models. Traditional word embeddings like Word2Vec and GloVe have been widely used due to their ability to capture semantic relationships between words based on their co-occurrence patterns in large corpora[34], [33]. These embeddings are particularly effective in supervised learning contexts, where they facilitate the classification of emotions into categories such as positive, negative, or neutral[34]. However, they often fall short in capturing the contextual nuances of words, which has led to the development of more advanced techniques like contextualized embeddings, such as BERT and RoBERTa, which consider the context of each word within a sentence [40]. These models have shown superior performance in sentiment analysis tasks, especially when integrated with deep learning architectures like CNNs and Bi-LSTMs, as they can grasp complex contextual nuances and improve classification accuracy[38], [39]. Moreover, recent advancements have introduced graph embedding techniques, which represent text as graphs to capture semantic and syntactic information more effectively, outperforming traditional word embeddings in certain sentiment analysis tasks[34]. Additionally, enhancements like Positional Embedding, which integrates word order into GloVe embeddings, have demonstrated improved accuracy in sentiment classification, highlighting the ongoing evolution and refinement of word embedding techniques[37]. Despite these advancements, challenges remain, such as handling out-of-vocabulary words and sentiment shifts over time, which continue to drive research in optimizing word embeddings for sentiment analysis[34], [36]. Overall, the choice of word embedding technique can significantly impact the performance of sentiment analysis models, necessitating a careful consideration of factors such as corpus size, content, and the specific requirements of the task at hand[36].Also, Word embedding representation for sentiment analysis utilizes dense low-dimensional space vectors to capture semantic similarities between words, overcoming the limitations of one-hot encoding. These embeddings enhance the performance of machine learning algorithms by providing richer contextual information. [35] explores various word representation techniques to determine which yields the highest performance in sentiment analysis when applied to both traditional and deep learning models, addressing the need for effective representations in extracting abstract information from text data.

D. Challenges in Sentiment Analysis for Arabic Dialects

Sentiment analysis for Arabic dialects faces several significant challenges primarily due to the complexity of the Arabic language. These challenges include morphological intricacies, limited resources, and the existence of various dialects, which complicate the extraction of sentiment from text [14]. Additionally, the informal nature of online Arabic content often disregards grammatical and spelling rules, making it difficult for traditional sentiment analysis tools to function effectively. The ambiguity in polarity, implicit sentiments, and the presence of sarcasm further complicate the analysis [14], [4]. Moreover, the short and noisy text typical of social media platforms, such as tweets, poses additional hurdles for feature extraction and accurate classification [15]. To address these issues, there is a pressing need for tailored methodologies and robust lexicons specifically designed for Arabic dialects [16], [66].

Word embedding representation plays a crucial role in Arabic sentiment analysis, as it helps capture the semantic nuances of the language, which is characterized by complex morphology and diverse dialects. Several studies have explored different word embedding techniques to enhance sentiment analysis in Arabic. AraBERT, a transformer-based model specifically designed for Arabic, has been widely used due to its ability to capture rich contextual information. It has been integrated with various neural network architectures such as Multi-channel Convolutional Neural Networks (MCNN), Bidirectional Gated Recurrent Units (BiGRU), and Long Short-Term Memory (LSTM) to improve sentiment classification accuracy[41], [42], [43]. Additionally, classical word embeddings like Word2Vec and fastText have been employed, with fastText showing superior performance in certain datasets[44]. Comparative analyses have demonstrated that contextualized embeddings like AraBERT generally outperform classical embeddings, achieving higher accuracy, precision, and recall[45]. The integration of these embeddings with deep learning models such as CNNs, LSTMs, and BiLSTMs has been shown to significantly enhance the performance of sentiment analysis systems, with BiLSTM often outperforming CNN in larger datasets[45]. These findings underscore the importance of selecting appropriate word embeddings and neural network architectures to effectively address the challenges posed by Arabic sentiment analysis.

As a result, most natural language processing (NLP) work in Arabic has traditionally focused on Modern Standard Arabic, as it is more standardized and has better-developed resources, such as large corpora and annotated datasets. However, this focus on MSA often neglects the practical, everyday language usage in Arabic-speaking communities, where dialects like Moroccan Darija dominate spoken interactions. Performing sentiment analysis on these dialects introduces additional challenges because they are primarily oral languages with little to no formal written structure, and they often vary significantly even within the same country.

For sentiment analysis tasks, the disparity between MSA and dialects poses considerable difficulties for NLP models. Models trained exclusively on MSA data struggle to adapt to dialectal texts, which are filled with colloquial expressions, informal grammar, and region-specific vocabulary. This makes it essential to either fine-tune pre-trained models on dialect-specific datasets or develop models that are specifically tailored to the linguistic characteristics of each dialect. In the case of Moroccan Darija, the language's informal nature and lack of standardized spelling make it even more challenging to process, requiring specialized approaches to handle these variations effectively.

IV. Experimental evaluation

This study involved the utilization of Moroccan Darija data, which we trained using four Arabic language models. Among these, three were pre-trained on Arabic dialect data from various countries, while one was specifically pre-trained on Moroccan dialect.

A. Experimental configuration

The computational framework for our proposed approach was implemented using Python 3.10 on a system running Windows 10. This system is equipped with an Intel(R)

Xeon(R) Gold 6230 CPU, which operates at 2.10 GHz across 160 cores and 4 sockets, and includes 64GB of RAM. This hardware configuration ensures a robust and reliable environment, providing consistent performance that supports the reproducibility of our research and facilitates further investigations.

B. Librairies

To enhance the reproducibility of our framework, we incorporated several well-established libraries, each chosen for its reliability and broad acceptance within the research community. Specifically, we employed the sklearn library for the execution of various machine learning algorithms, which is renowned for its comprehensive collection of tools. Additionally, we utilized ModAL to integrate Active Machine Learning functionalities effectively, and the pandas library was used for efficient data manipulation tasks. We also integrated Hugging Face's transformers library to leverage state-of-the-art transformer models for natural language processing tasks. The selection of these libraries ensures that our methodologies can be easily adopted and replicated by other researchers, thereby fostering an environment of collaborative and verifiable scientific inquiry.

C. Dataset

For this study, we utilized the large Moroccan dialect database developed by the LIPIM laboratory comprises 20,000 sentences written in Moroccan Darija, sourced from a variety of digital platforms, including social media, blogs, and forums. These platforms provide a wealth of informal language data, which is critical for understanding how Darija is used in everyday communication, especially in its written form. Given that Moroccan Darija is primarily a spoken dialect, capturing written instances of the language is particularly valuable for sentiment analysis, as it reflects the natural linguistic behaviors of its speakers. The dataset has been meticulously labeled for sentiment, categorizing each sentence as positive or negative. This labeling is essential for training the transformer models to recognize emotional tone and opinion within Darija texts. As Darija lacks a formalized writing system, considerable preprocessing was necessary to prepare the dataset for model training. Preprocessing involved cleaning the text by removing non-Darija words, which included foreign words that were not integrated into the dialect and irrelevant symbols or characters.

By integrating texts written in both Arabic script and Arabizi, this dataset captures the full range of how Moroccan Darija is expressed in written form. This diversity ensures that the models trained on this dataset are robust and capable of handling the various ways in which Darija speakers write their language, whether formally in Arabic script or informally in Latin characters. This approach is particularly valuable for sentiment analysis, as it allows the models to better understand and interpret sentiments expressed across different writing styles.

D. Preprocessing and Tokenization

To ensure compatibility with the transformer models, the text data was tokenized using appropriate tokenizers. For Arabic text, we used the BERT tokenizer designed for Arabic, while for Arabizi, we adapted tokenizers to handle Latin characters. The following preprocessing steps were applied:

- Tokenization: Text was split into subword tokens using the respective tokenizer.
- Padding and Truncation: Sentences were padded to a length of 128 tokens to ensure uniform input size.
- Attention Masks: Generated attention masks to help the model focus on non-padding tokens.

E. Evaluation Metrics

The performance of each model was evaluated using well-established classification metrics, which are widely utilized in natural language processing (NLP) tasks to assess the accuracy and reliability of machine learning models. These metrics offer a comprehensive understanding of how well each model can perform sentiment analysis, ensuring a balanced evaluation of their strengths and weaknesses. The key metrics used in this evaluation are as follows:

- Accuracy: This metric represents the overall effectiveness of the model by measuring the percentage of correct predictions made. It is calculated as the ratio of correctly predicted sentiments (positive, negative, and neutral) to the total number of predictions. Accuracy provides a general view of the model's performance, but it does not offer insights into the performance of individual sentiment categories, especially in cases of class imbalance.
  It is calculated by: Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- Precision: Precision measures the accuracy of positive sentiment predictions. Specifically, it is the ratio of correctly predicted positive sentiments to the total number of positive predictions made by the model. High precision indicates that when the model predicts a positive sentiment, it is likely to be correct. Precision is particularly useful in scenarios where false positives (incorrectly predicting positive sentiment) need to be minimized.
  It is expressed as: Precision $= \frac{TP}{TP+FP}$
- Recall: Also known as sensitivity or true positive rate, recall evaluates the model's ability to correctly identify all instances of positive sentiment in the dataset. It is the ratio of correctly predicted positive sentiments to the total actual positive sentiments present in the dataset. A high recall indicates that the model can successfully capture most of the positive sentiments, though it may come at the expense of incorrectly classifying some neutral or negative sentiments as positive.
  It is formulated as: Recall $= \frac{TP}{TP+FN}$
- F1-Score: The F1-Score is the harmonic mean of precision and recall, offering a single metric that balances the trade-off between these two measures. By combining both precision and recall into one score, the F1-Score helps to provide a more comprehensive evaluation of the model's performance, especially in cases where there is an imbalance between the sentiment classes or where one class is harder to predict than others. It is particularly valuable when a balance between false positives and false negatives is

required, as it takes both into account.

The F1-score is calculated as follows: $F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$

These metrics together provide a detailed and nuanced assessment of each model's performance in sentiment analysis, helping to identify which model performs best not only in terms of overall accuracy but also in its ability to correctly identify and classify sentiment across various categories. By using these metrics, we can ensure that the evaluation captures both the model's strengths in prediction accuracy and its potential limitations, such as its handling of false positives and false negatives.

V. Comparative Study: Methodology and Results

This section presents a comparative study of transformer models for sentiment analysis in Moroccan Darija. We describe the transformers workflow 1 and the models evaluated, and also the word embedding representations used, outline the metrics used for assessing performance, and provide a detailed discussion of the results. Additionally, we examine the challenges encountered during the study and propose directions for future research to improve sentiment analysis in dialectal Arabic.

A. *Transfomers workflow*

- Data Collection The process starts with gathering a dataset specifically labeled for sentiment analysis. This dataset forms the basis of the model's training and evaluation. In sentiment analysis, the data should ideally reflect the language or dialect's characteristics, with samples categorized into sentiment labels like positive, negative, or neutral. When dealing with dialect-specific tasks, such as Moroccan Darija, it is important that the data is representative of the dialect's informal language use, varied expressions, and potential code-switching.
- Data Pre-processing Once the data is collected, it goes through a comprehensive pre-processing phase. This involves several steps to prepare the text for model input. Cleaning is the first step, where non-essential elements like irrelevant characters and stopwords are removed, ensuring that only meaningful content remains. Following this, tokenization splits the text into smaller units or subwords, which helps capture the complexities of the language and improves the model's ability to understand nuanced expressions. Normalization is then applied to standardize variations in spelling and manage different scripts, especially in cases where languages may be written using multiple alphabets, like Arabic and Latin in the case of Arabizi. Padding and truncation are used to make all text sequences uniform in length, which is necessary for batch processing in transformer models.
- Model Selection The next step involves selecting an appropriate pre-trained transformer model based on the specific requirements of the task. General models like BERT or RoBERTa are available for broad applications, while dialect-specific models such as DarijaBERT or MARBERT are chosen for tasks involving regional dialects. The choice of model depends on factors like the language variety, data availability, and the desired level of specificity in sentiment analysis.
- Fine-tuning After selecting a model, it is fine-tuned on the labeled dataset. Fine-tuning involves adapting the general pre-trained model to the specific sentiment analysis task by optimizing it on the collected data. During this step, hyperparameters such as learning rate, batch size, and the number of epochs are adjusted to achieve the best possible model performance. Fine-tuning is particularly important for tasks involving specific dialects, as it helps the model learn the subtle linguistic patterns and unique vocabulary of the target language.
- Model Evaluation Following fine-tuning, the model's performance is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics offer a comprehensive view of the model's effectiveness in correctly classifying sentiments. Evaluating the model allows for assessing its strengths and identifying any areas where it may need improvement, ensuring that it meets the required accuracy for real-world applications.
- Error Analysis and Iteration Once the model has been evaluated, an error analysis is conducted to examine instances where the model misclassified sentiments. This step provides valuable insights into any recurring issues, such as misinterpretation of certain phrases or confusion between sentiment classes. Based on these findings, adjustments are made to the pre-processing steps or fine-tuning process. The model may go through several iterations of training and adjustment to enhance its accuracy and reliability.
- Deployment and Testing After achieving satisfactory results, the model is deployed for real-world use. During deployment, continuous testing and monitoring are essential to ensure the model maintains its performance over time. This step allows for adaptive modifications as new data becomes available or as the language evolves. Real-world testing ensures that the model can handle diverse input while maintaining consistent sentiment analysis accuracy.
- Future Enhancements The final step in the workflow focuses on future improvements. As language use changes over time, new data can be collected periodically to retrain or further refine the model, ensuring it remains accurate and effective. Additionally, future research may explore advanced architectures or additional layers of fine-tuning to improve the model's adaptability and performance in sentiment analysis tasks involving dialectal language.

B. *Language Models evaluated*

In this study, we fine-tuned six different transformer models specifically for sentiment analysis on the Moroccan Darija dataset. Each of these models brings unique features and pre-training strategies, making them well-suited for different aspects of natural language processing tasks, including sentiment classification. The models fine-tuned on the Darija dataset are as follows: BERT Multilingual, DarijaBERT, DarijaBERT Arabizi, MARBERT, AraBERT,

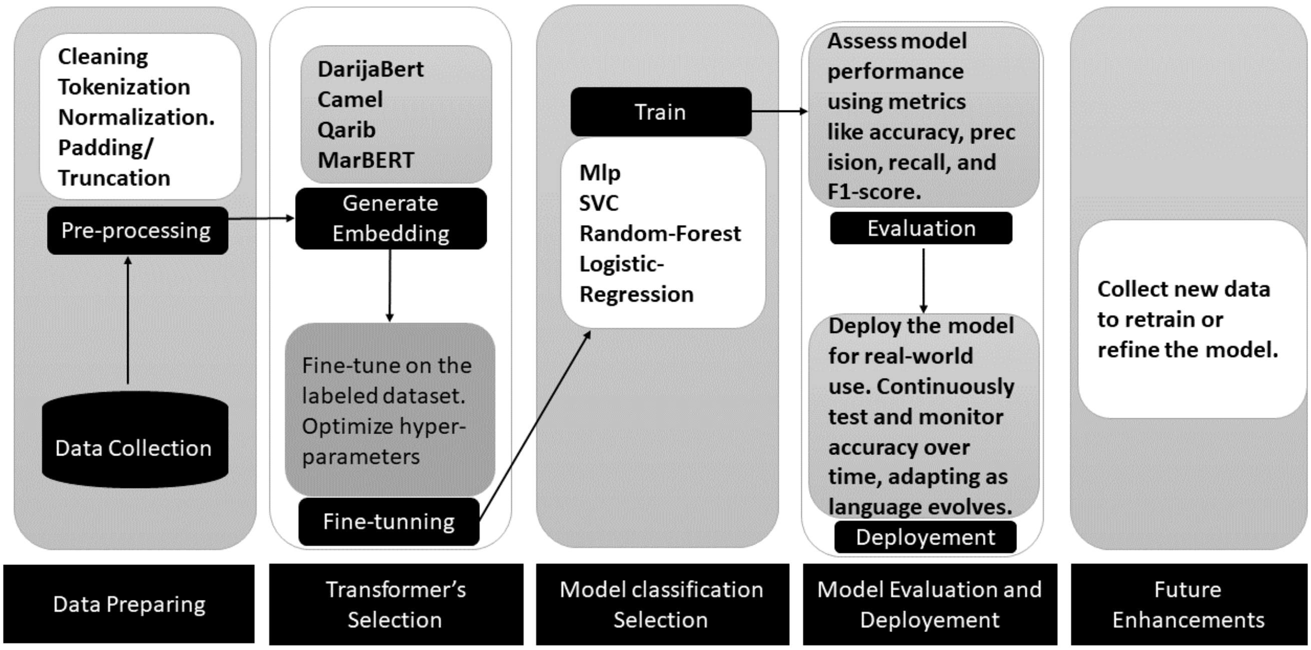
Cleaning
Tokenization
Normalization.
Padding/
Truncation
Pre-processing
Data Collection
DarijaBert
Camel
Qarib
MarBERT
Generate Embedding
Fine-tune on the labeled dataset. Optimize hyper-parameters
Fine-tunning
Train
Mlp
SVC
Random-Forest
Logistic-Regression
Assess model performance using metrics like accuracy, prec ision, recall, and F1-score.
Evaluation
Deploy the model for real-world use. Continuously test and monitor accuracy over time, adapting as language evolves.
Deployement
Collect new data to retrain or refine the model.
Data Preparing
Transformer's Selection
Model classification Selection
Model Evaluation and Deployement
Future Enhancements

Fig. 1. Transformer Model Workflow for Effective Sentiment Analysis.

CAMELBERT. Each model, depending on its architecture and pre-training data, offers distinct advantages and faces particular challenges when it comes to analyzing sentiment in a low-resource dialect like Darija.

1) BERT Multilingual:: This version of BERT has been pre-trained on 104 languages, including Arabic, which allows it to support multilingual tasks. Its broad linguistic range makes it useful for analyzing texts in multiple languages without requiring separate models for each. However, because it is trained on a generalized dataset, BERT Multilingual may struggle to capture the intricacies and specificities of particular dialects like Darija, where unique vocabulary and syntactic patterns are prevalent. While it offers a foundation for sentiment analysis in Arabic, its performance might be limited when applied to highly informal or region-specific dialects that deviate significantly from MSA.

2) DarijaBERT: A transformer model specifically tailored to the Moroccan Darija dialect, it has been pre-trained on a dataset that includes the unique grammatical structures and lexicon of Darija. This model is designed to excel in tasks specific to Darija, capturing the nuances that are often missed by models trained on more formal versions of Arabic or on general multilingual datasets. By focusing exclusively on this dialect, DarijaBERT is better equipped to handle sentiment analysis in texts that reflect the informal and highly variable nature of spoken Darija, which is often quite different from formal written Arabic.

3) DarijaBERT Arabizi: Similar to DarijaBERT, this model has been trained specifically on Darija, but with a key difference—it has been pre-trained on texts written in Arabizi, a form of writing where Arabic dialects, including Darija, are transcribed using the Latin alphabet. Arabizi is commonly used in informal digital communication, particularly on social media platforms. The challenge with Arabizi lies in the lack of standardization in spelling and grammar, which makes it difficult for traditional models to interpret. However, DarijaBERT Arabizi is designed to understand this writing style and is well-suited for tasks like sentiment analysis in informal online texts.

4) MARBERT: This model was developed with the goal of addressing sentiment analysis across multiple forms of Arabic, including both MSA and a variety of dialects. Pre-trained on a large dataset that includes texts from social media and other informal sources, MARBERT is specifically designed to handle the blending of formal and colloquial Arabic. While it is versatile enough to work with Moroccan Darija, its broader focus on multiple dialects means it may not be as specialized for Darija as models like DarijaBERT. Nevertheless, MARBERT remains a powerful option for tasks that involve a mix of formal Arabic and regional dialects.

5) AraBERT: AraBERT is a widely used model pre-trained on MSA, making it well-suited for a range of general Arabic NLP tasks such as sentiment analysis, text classification, and question-answering. However, since AraBERT is trained on MSA, which is more formal and structured compared to regional dialects, it may not perform as well when applied to informal texts or specific dialects like Moroccan Darija. The model's lack of exposure to dialectal variations can limit its ability to capture the linguistic subtleties present in spoken or informal Arabic.

6) CAMEL BERT: Designed to focus on Arabic dialects, CAMEL BERT has been trained across several regional variations, including Levantine, Gulf, and Egyptian dialects. While this makes CAMEL BERT a versatile model for dialectal Arabic, it lacks a specific focus on Moroccan Darija, which has its own distinct vocabulary and grammar. This model is therefore best suited for tasks that involve multiple Arabic dialects, though it may not perform as well on Darija-specific tasks compared to models like DarijaBERT or MARBERT.

7) Benchmarking Language Models for Sentiment Analysis in Moroccan Darija: To evaluate the performance of the fine-tuned transformer models on the task of sentiment analysis in Moroccan Darija, we employed several standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics offer a comprehensive assessment of each model's ability to correctly classify sentiment into positive or negative categories. The table below summarizes the performance of six transformer models: DarijaBERT, DarijaBERT Arabizi, MARBERT, AraBERT, BERT Multilingual, and CAMEL BERT, highlighting their effectiveness in this specific context

TABLE I
PERFORMANCE OF TRANSFORMER MODELS ON DARIJA SENTIMENT ANALYSIS

Model	Accuracy	Precision	Recall	F1-Score
DarijaBERT	88.2%	89.1%	87.5%	88.3%
DarijaBERT Arabizi	83.5%	84.0%	83.0%	83.5%
MARBERT	85.6%	86.4%	85.2%	85.8%
AraBERT	81.4%	82.2%	80.7%	81.4%
BERT Multilingual	79.3%	79.7%	78.6%	79.1%
CAMEL BERT	80.1%	80.7%	79.5%	80.0%

The performance results of the fine-tuned models reveal critical insights into the importance of dialect-specific pre-training for sentiment analysis in Moroccan Darija. Among the models evaluated, DarijaBERT clearly emerged as the top performer, demonstrating that pre-training on data specific to the Darija dialect is essential for accurately capturing its unique linguistic features. The model's superior performance highlights the need for targeted approaches when dealing with regional dialects, where vocabulary, syntax, and cultural expressions differ significantly from more standardized forms of Arabic, such as Modern Standard Arabic (MSA).

DarijaBERT Arabizi also delivered strong results, further underscoring the importance of model specialization. By being specifically trained on Darija texts written in Latin characters (Arabizi), this model is uniquely equipped to process and analyze informal, digitally-mediated texts—particularly common in social media and online forums. The ability to handle the complex, non-standardized nature of Arabizi writing suggests that fine-tuning models to accommodate different writing systems within the same dialect can lead to substantial improvements in performance.

MARBERT, which was pre-trained on both MSA and various Arabic dialects, also showed promising results but fell slightly short of DarijaBERT. The model's broader focus on multiple dialects may have diluted its ability to fully grasp the specific nuances of Moroccan Darija, indicating that while MARBERT is versatile, it does not reach the same level of precision as models fine-tuned exclusively for Darija.

Conversely, models like AraBERT and BERT Multilingual encountered difficulties in dealing with the complexities of Darija. Both of these models were pre-trained on MSA or a wide range of multilingual texts, which left them less effective at capturing the informal and regionally specific characteristics of the Moroccan dialect. The results suggest that while these models are powerful for general Arabic NLP tasks, they struggle with the subtleties and idiosyncrasies of dialectal Arabic, particularly when dealing with highly informal or culturally specific language use.

Lastly, CAMEL BERT, despite being designed to handle Arabic dialects, did not outperform the models fine-tuned exclusively for Moroccan Darija. This finding reinforces the notion that pre-training models on data that directly reflects the target dialect is critical for achieving the best possible performance. While CAMEL BERT's broader training on multiple dialects makes it adaptable, it is not as effective as models that have been specifically tailored to a single dialect, such as DarijaBERT.

Overall, these results emphasize the need for specialized pre-training on dialect-specific data when performing sentiment analysis in Moroccan Darija, as models trained on more generalized or formal Arabic data struggle to capture the nuanced linguistic patterns and informal expressions prevalent in the dialect.

8) Visual Performance Analysis of Language Models: The graphical representation 2 of performance metrics provides a clear and concise comparison of the evaluated models, emphasizing the importance of pre-training tailored to Moroccan Darija for sentiment analysis. DarijaBERT outperforms all other models across all metrics, reinforcing its ability to capture the unique linguistic and cultural intricacies of Moroccan Darija. This highlights the impact of dialect-specific pre-training on improving model performance in dialectal NLP tasks. DarijaBERT Arabizi also performs robustly, demonstrating its specialization for handling informal Latin-scripted Darija text, a prevalent form of communication in digital and social media platforms. While models like MARBERT and CAMEL BERT exhibit moderate performance due to their broader training on multiple dialects, they fall short of achieving the precision offered by DarijaBERT.

The underperformance of AraBERT and BERT Multilingual further illustrates the limitations of generalized models when applied to highly informal and regionally specific dialects. This graph underscores the necessity of leveraging dialect-specific and domain-relevant embeddings to maximize accuracy, precision, recall, and F1 score in sentiment classification tasks.

C. Word Embedding representation

In this subsection, we introduce two widely recognized word embedding techniques, fastText and Word2Vec, to benchmark their performance against embeddings generated by advanced language models like DarijaBERT. These methods were chosen due to their proven efficiency in various natural language processing tasks, particularly in low-resource and dialectal contexts. fastText is known for its ability to capture subword information, making it highly effective for morphologically rich languages like Moroccan Darija. Similarly, Word2Vec has demonstrated robust performance in representing semantic relationships between words through its continuous bag-of-words (CBOW)

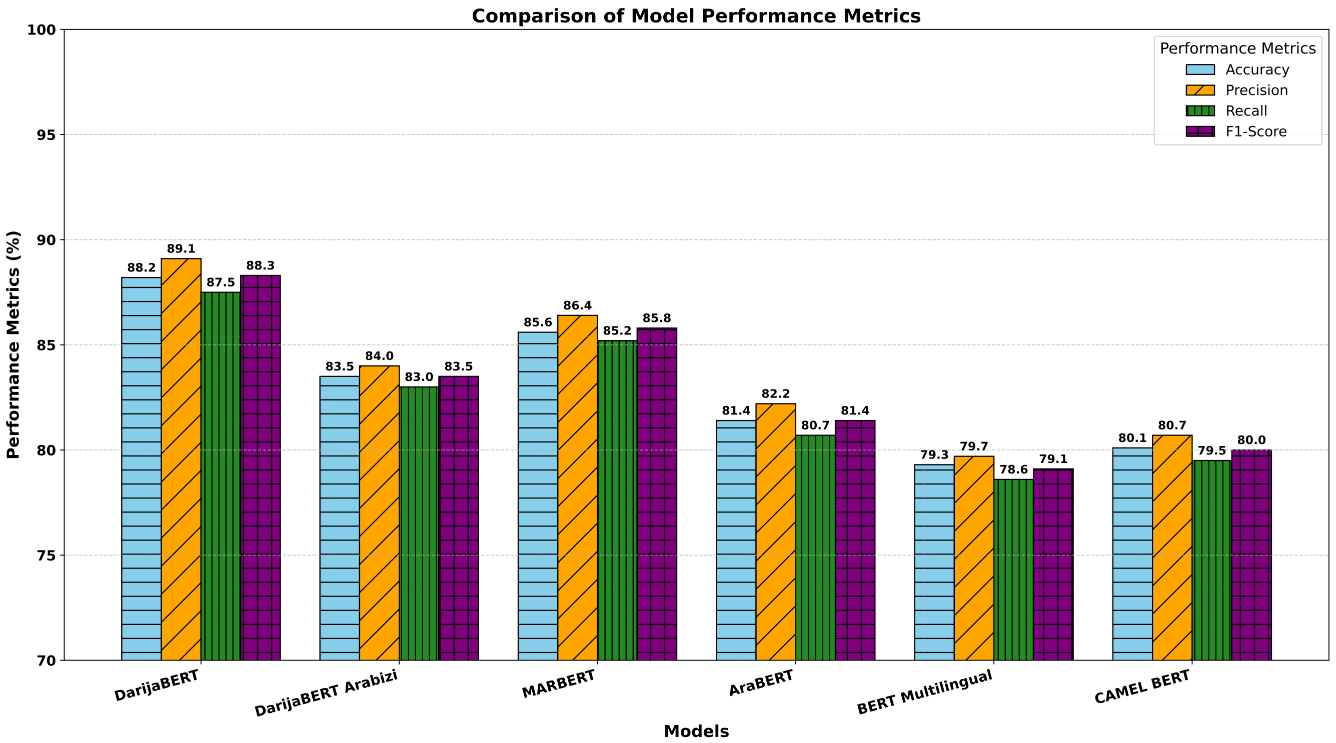

Fig. 2. Evaluating language models performance.

and skip-gram architectures. By including these methods, we aim to provide a comprehensive evaluation that highlights the strengths and limitations of traditional static embeddings compared to modern transformer-based embeddings. This approach ensures a balanced and insightful comparison for sentiment analysis in Moroccan Darija.

1) Fasttext: FastText is an efficient open-source library developed by Facebook AI, designed for natural language processing (NLP) tasks, particularly for word embeddings and text classification. It excels in sentiment analysis due to its ability to generate high-quality word representations and perform text classification rapidly.

Sentiment analysis using FastText word representation has been explored across various domains, demonstrating its versatility and effectiveness. It is known for its ability to handle out-of-vocabulary words and capture subword information and is particularly beneficial in sentiment analysis tasks. In educational contexts, combining FastText with BERT has been shown to effectively address issues like polysemy and new internet words, providing precise sentiment analysis of educational texts, such as those from online ideological and political courses[46]. In the realm of Arabic e-commerce reviews, FastText has been used alongside GloVe in a hybrid deep learning model, significantly enhancing sentiment prediction accuracy, as evidenced by a performance margin of 0.9112 over standard deep learning models[47]. Additionally, FastText has been combined with AraVec for sentiment analysis, achieving superior accuracy with machine learning algorithms like NuSVC, particularly in social media contexts where user-generated content is abundant[48]. On Twitter, FastText has been employed for aspect-based sentiment analysis, addressing vocabulary mismatch issues and improving model performance significantly, as seen in the analysis of Telkomsel's product reviews[49]. Furthermore, FastText's effectiveness in Twitter sentiment analysis has been validated through comparative studies, where it outperformed other word embeddings like Word2Vec and GloVe when used with classifiers such as NuSVC[50]. These studies collectively highlight FastText's robust application in sentiment analysis across different languages and platforms, underscoring its adaptability and precision in handling diverse textual data.

2) Word2vec: Sentiment analysis using Word2Vec word representation techniques has been explored across various studies, highlighting its effectiveness and adaptability in different contexts. Word2Vec, known for its simplicity and efficiency, remains a popular choice for sentiment analysis, especially in resource-constrained environments. Liu's study demonstrates that Word2Vec embeddings, particularly multichannel embeddings combining static and non-static representations, perform well across neural network architectures like CNNs, LSTMs, and GRUs, achieving a balance between efficiency and accuracy[51]. Irianto et al. applied Word2Vec with the K-Nearest Neighbors (KNN) algorithm to analyze customer reviews, achieving a high F1-score of 91.98%, which underscores the method's effectiveness in capturing sentiment patterns in customer feedback[52]. Kumar et al. emphasize that Word2Vec, alongside other word embedding techniques, facilitates the transformation of text into vectors, enabling the application of both machine learning and deep learning methods for sentiment classification, with deep learning approaches generally outperforming traditional machine learning in accuracy. Ayar et al. introduce Word2HyperVec, a novel framework combining Word2Vec with Hyperdimensional Computing, which offers a lightweight and efficient solution for sentiment analysis on edge devices, highlighting the adaptability of Word2Vec in emerging computational paradigms[53]. Vinluan's research further refines Word2Vec by improving the initialization algorithm of word vectors, resulting in enhanced performance for sentiment polarity classification, demonstrating the ongoing evolution and optimization of Word2Vec techniques in sentiment analysis[54]. Collectively, these studies illustrate the versatility and continued relevance of Word2Vec in sentiment analysis across various applications and computational settings.

3) FastText vs Word2Vec: Key Differences: FastText and Word2Vec are two prominent word embedding techniques used in natural language processing, particularly for tasks like sentiment analysis. Here's a comparative analysis of these methods based on their features, performance, and suitability for sentiment analysis.

- Architecture and Approach:
  Word2Vec: Developed by Google, Word2Vec relies on two main architectures: Continuous Bag of Words (CBOW) and Skip-gram. These methods learn word representations by predicting a word based on its context (CBOW) or the context based on a word (Skip-gram). While Word2Vec effectively captures semantic relationships, it has a notable limitation in handling out-of-vocabulary (OOV) words, as it generates embeddings only for words encountered in the training corpus. FastText: Created by Facebook, FastText extends the capabilities of Word2Vec by representing words as combinations of character-level n-grams. This approach enables FastText to handle OOV words by breaking them into subword units, making it particularly well-suited for morphologically rich languages and for dealing with misspellings or rare terms.
- Performance in Sentiment Analysis Studies have demonstrated that FastText consistently outperforms Word2Vec in sentiment analysis tasks. FastText's ability to leverage subword information allows it to achieve higher accuracy, particularly when working with languages that exhibit rich morphology or variability in word forms. For example, when tested on sentiment classification tasks using models like Random Forest or Support Vector Machines, FastText has shown superior performance compared to Word2Vec, highlighting its robustness in capturing nuanced linguistic patterns.
- Generalization and Robustness
  Generalization: FastText's reliance on character-level n-grams allows it to generalize better to unseen words, an essential feature in sentiment analysis where new terms, slang, or domain-specific vocabulary frequently appear. In contrast, Word2Vec struggles with providing meaningful embeddings for such cases, as it depends on a fixed vocabulary derived from the training data. Robustness: FastText is particularly robust against linguistic variations, such as misspellings and morphological changes, making it ideal for complex dialects like Moroccan Darija. Word2Vec, while effective for standard language patterns, is less adaptable to such complexities.

In Conclusion,FastText generally provides superior performance for sentiment analysis compared to Word2Vec due to its innovative use of subword information and its ability to handle OOV challenges effectively. While Word2Vec offers a strong baseline for understanding word semantics, FastText's versatility and robustness make it a preferred choice for applications involving diverse and linguistically complex data.

D. Results and Discussion

1) Strategic Classifier Selection for Moroccan Darija Data: The selection of Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), Logistic Regression, and Random Forest for sentiment analysis on Moroccan Darija data is driven by their unique strengths in addressing the challenges posed by dialectal language processing. MLP, as a deep learning model, is adept at capturing complex and non-linear relationships, making it ideal for leveraging advanced embeddings like DarijaBERT, which encapsulate rich contextual and semantic nuances of Moroccan Darija. Its ability to learn from high-dimensional representations ensures that no intricate pattern within the data is overlooked.

SVM was chosen for its efficiency in handling high-dimensional feature spaces and its effectiveness in generating robust decision boundaries. This makes SVM particularly suitable for Moroccan Darija sentiment analysis, where subtle differences in linguistic expressions play a critical role in classification. Its versatility in working with both small and moderately sized datasets ensures consistent and reliable performance.

Logistic Regression provides a strong baseline due to its simplicity and interpretability, offering a clear understanding of how the features contribute to classification outcomes. This model is computationally efficient and effective for binary and multi-class sentiment classification, making it an indispensable tool for comparative evaluations.

Random Forest, as an ensemble learning method, excels in capturing feature interactions and reducing the risk of overfitting, which is crucial in datasets with linguistic variability like Moroccan Darija. Its ability to aggregate predictions from multiple decision trees ensures robust performance, particularly when dealing with diverse text patterns and morphologically rich language structures. By combining these classifiers, this study achieves a balanced evaluation of sentiment analysis performance, leveraging their complementary strengths to address the unique challenges of Moroccan Darija text processing

2) DarijaBert: The performance metrics presented in Table II demonstrate the effectiveness of DarijaBERT embeddings across various machine learning classifiers for sentiment analysis in Moroccan Darija. Among the classifiers evaluated, the Multi-Layer Perceptron (MLP) achieved the highest accuracy, F1 score, precision, and recall, with all

TABLE II
Model Performance Metrics of DarijaBert classication

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	95.35%	95.35%	95.35%	95.35%
SVM	95.07%	95.07%	95.07%	95.07%
Random Forest	91.90%	91.90%	91.90%	91.90%
MLP	96.00%	95.99%	96.00%	96.00%

metrics exceeding 96%. This result underscores the ability of neural network-based models to effectively leverage the contextual richness of DarijaBERT embeddings, particularly in capturing the complex, non-linear relationships inherent in dialectal text. Logistic Regression and Support Vector Machines (SVM) also performed exceptionally well, with accuracies of 95.35% and 95.07%, respectively. These results highlight that even traditional classifiers can produce competitive results when powered by high-quality embeddings, with Logistic Regression slightly outperforming SVM, likely due to its efficiency in handling linearly separable data. In comparison, the Random Forest classifier achieved a slightly lower accuracy of 91.90%, suggesting that ensemble tree-based models may be less effective in fully utilizing the dense, high-dimensional representations generated by DarijaBERT. Notably, the consistency across all metrics for each classifier—accuracy, F1 score, precision, and recall—indicates a balanced predictive performance across sentiment categories, with minimal trade-offs. These findings confirm the robustness and versatility of DarijaBERT embeddings in supporting sentiment classification tasks, whether through complex neural networks or simpler, computationally efficient models. This versatility makes DarijaBERT a valuable tool for sentiment analysis in low-resource, dialectal contexts like Moroccan Darija.

3) Fasttext: The performance metrics presented in Table III highlight the effectiveness of fastText embeddings in sentiment analysis when combined with various machine learning classifiers. Among the models evaluated, Random Forest achieved the highest accuracy and F1 score at 76.49% and 76.43%, respectively, indicating its strong ability to capitalize on the word-level representations provided by fastText. The Multi-Layer Perceptron (MLP) followed closely, with slightly lower accuracy and F1 scores but a marginally higher precision of 76.80%, showcasing its capability to model complex patterns in data.

TABLE III
MODEL PERFORMANCE METRICS OF FASTTEXT

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	73.34%	73.13%	73.86%	73.34%
SVM	75.02%	74.68%	76.12%	75.02%
Random Forest	76.49%	76.43%	76.65%	76.49%
MLP	76.32%	76.17%	76.80%	76.32%

Support Vector Machines (SVM) and Logistic Regression demonstrated competitive performance, with accuracies of 75.02% and 73.34%, respectively. These results suggest that traditional classifiers can effectively utilize fastText embeddings, although they may not fully exploit the intricate relationships captured by these embeddings compared to ensemble or neural network-based models. Notably, SVM's superior precision at 76.12% reflects its strength in boundary-based classification, particularly when embeddings provide robust contextual information.

Despite these promising results, the overall performance of fastText embeddings is modest compared to modern transformer-based models, as evidenced by the metrics presented here. This outcome aligns with the inherent limitations of static embeddings, such as their inability to account for polysemy or contextual variability. Nevertheless, fastText's ability to handle out-of-vocabulary (OOV) words through its character-level n-gram approach ensures reasonable robustness, especially in morphologically rich languages like Moroccan Darija. These findings underline the utility of fastText embeddings as a lightweight, computationally efficient alternative for sentiment analysis, particularly in resource-constrained environments.

4) Word2vec: The performance metrics presented in Table IV provide insights into the effectiveness of Word2Vec embeddings when paired with various machine learning classifiers for sentiment analysis tasks. Among the classifiers, the Multi-Layer Perceptron (MLP) achieved the highest accuracy and F1 score, with values of 76.37% and 76.19%, respectively. This result highlights the ability of MLP to leverage the contextual relationships captured by Word2Vec embeddings.

TABLE IV
MODEL PERFORMANCE METRICS OF WORD2VEC

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	73.42%	73.06%	74.40%	73.42%
SVM	73.34%	72.93%	74.50%	73.34%
Random Forest	74.52%	74.42%	74.76%	74.52%
MLP	76.37%	76.19%	76.96%	76.37%

Random Forest followed closely, achieving an accuracy of 74.52% and an F1 score of 74.42%. These metrics demonstrate the suitability of ensemble methods for utilizing the dense, fixed-dimensional embeddings generated by Word2Vec. Logistic Regression and Support Vector Machines (SVM) performed slightly lower, with accuracies of 73.42% and 73.34%, respectively. However, their precision scores, particularly SVM's 74.50% and Logistic Regression's 74.40%, indicate a strong ability to identify relevant patterns within the data.

While Word2Vec embeddings enable robust performance across classifiers, the overall metrics reveal limitations inherent to static embeddings. Unlike context-aware embeddings, Word2Vec does not adapt dynamically to the surrounding words, which may lead to reduced effectiveness in handling linguistic nuances and polysemy. Nevertheless, Word2Vec remains a reliable choice for sentiment analysis, offering a computationally efficient solution for tasks where the context of individual words remains consistent. These results highlight Word2Vec's utility as a foundational embedding technique, particularly in scenarios requiring moderate accuracy with minimal computational overhead.

5) Graphical Analysis of Model Performance : The F1 score comparison for Logistic Regression 3 demonstrates that while the model effectively utilizes embeddings for sentiment classification, there is a noticeable difference in performance across the three methods. DarijaBERT embeddings significantly outperform both FastText and Word2Vec, showcasing their ability to capture contextual nuances specific to Moroccan Darija. The lower scores for FastText and Word2Vec suggest limitations in handling the complexities of dialectal language, particularly in static embedding methods. Logistic Regression's linear approach benefits from the rich contextual information provided by DarijaBERT, leading to higher classification accuracy and balanced predictions.

The F1 score comparison for MLP 4 highlights its superior

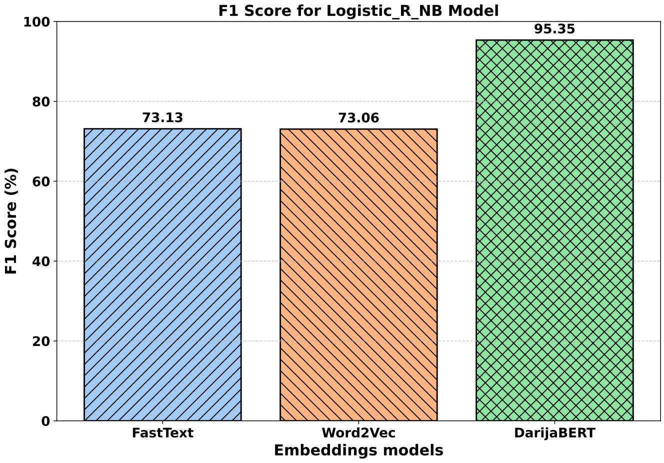

Fig. 3. F1 Score Comparison for Logistic Regression.

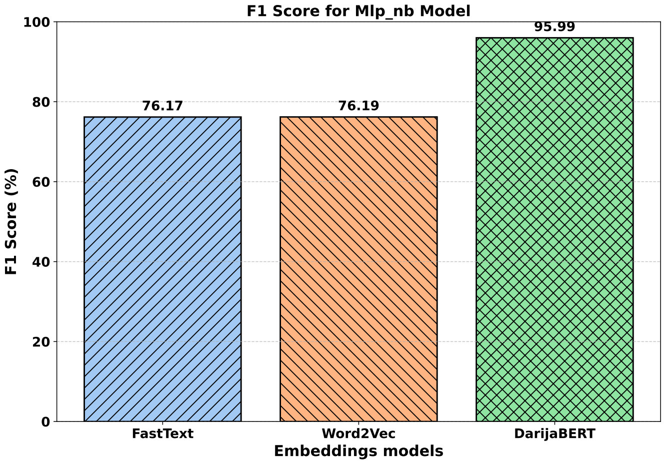

Fig. 4. F1 Score Comparison for MLP.

performance when paired with DarijaBERT embeddings. As a neural network-based classifier, MLP excels at leveraging high-dimensional representations, which explains its higher F1 score compared to other embedding methods. The results underscore the strength of DarijaBERT in providing comprehensive embeddings for sentiment analysis, while FastText and Word2Vec, although effective, lack the contextual depth required for optimal performance. This demonstrates the synergy between advanced embeddings and deep learning models like MLP in handling complex language patterns inherent in Moroccan Darija. The F1 score comparison for Random Forest 5 reveals a consistent yet slightly reduced performance compared to other classifiers. DarijaBERT embeddings lead to the highest F1 score, showcasing their ability to enhance decision-making in ensemble models. FastText and Word2Vec provide competitive results, but their static nature limits their adaptability to intricate language variations. Random Forest's ability to aggregate predictions from multiple decision trees benefits significantly from DarijaBERT's rich and dynamic embeddings, resulting in improved sentiment classification for Moroccan Darija, though it remains less effective than MLP in fully utilizing these embeddings.

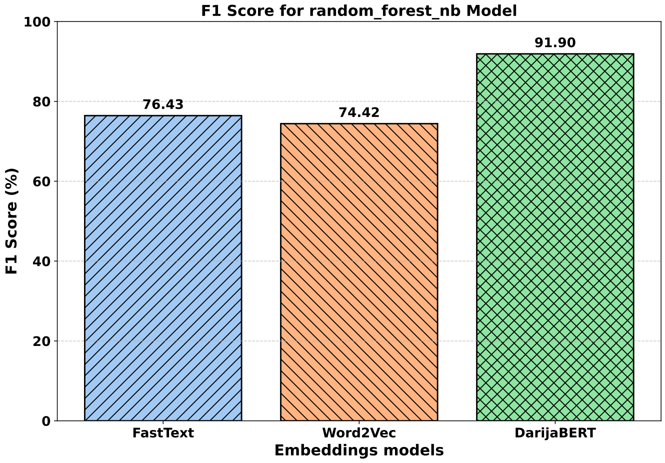

Fig. 5. F1 Score Comparison for Random Forest.

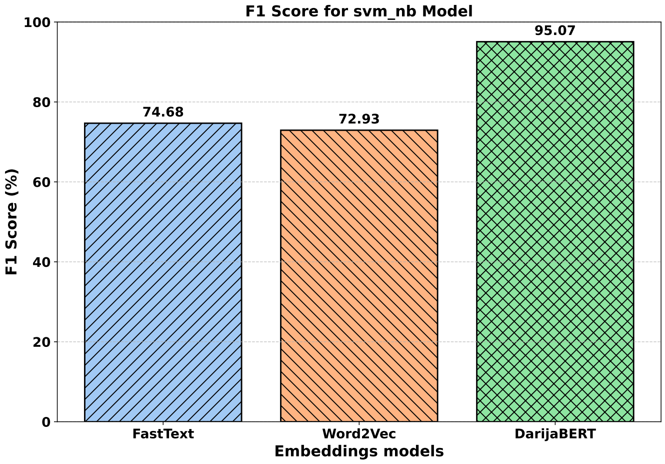

Fig. 6. F1 Score Comparison for SVM.

The F1 score comparison for SVM see figure 6 illustrates a clear advantage for DarijaBERT embeddings, which significantly outperform FastText and Word2Vec. SVM's strength in creating robust decision boundaries is maximized with the high-quality, context-aware embeddings of DarijaBERT. In contrast, the performance with FastText and Word2Vec suggests that static embeddings are less equipped to handle the subtle contextual cues required for accurate sentiment classification. These results emphasize the importance of using context-rich embeddings in tasks involving complex dialects like Moroccan Darija.

E. Performance Overview: Contextual vs Static Embeddings

The findings presented in figure 7 in this section underscore the superior performance of DarijaBERT embeddings compared to traditional embedding techniques such as FastText and Word2Vec across all evaluated machine learning classifiers see figure 7. This consistent outperformance demonstrates the value of leveraging contextualized embeddings specifically tailored for dialectal languages like Moroccan Darija. Unlike static embeddings, which struggle to capture the linguistic nuances and variability of the dialect, DarijaBERT effectively models the rich semantic and syntactic structures unique to the language, resulting in higher F1 scores and improved overall classification performance. The study's comprehensive approach, which combines multiple embedding methods and classifiers, provides a robust evaluation framework that highlights the strengths and limitations of each technique. By including widely used machine learning models such as Logistic Regression, SVM, Random Forest, and MLP, this research ensures a balanced comparison and validates the adaptability of DarijaBERT embeddings across diverse classification algorithms. The superior performance of DarijaBERT, particularly when paired with deep learning models like MLP, further reinforces the importance of integrating domain-specific embeddings for complex NLP tasks in under-resourced languages.

This comparative study not only demonstrates the practical advantages of using DarijaBERT for sentiment analysis but also sets a benchmark for future research in the field. By providing empirical evidence of the efficacy of contextualized embeddings in dialectal contexts, this work contributes valuable insights into the development of NLP tools for low-resource languages. The findings pave the way for more effective sentiment analysis systems and highlight the potential of transformer-based models to bridge the gap in linguistic representation for less-studied languages like Moroccan Darija.

Overall, this study showcases the importance of tailored language models in achieving state-of-the-art results for sentiment classification and provides a solid foundation for future advancements in the application of NLP to dialectal and low-resource languages.

TABLE V
CLASSIFICATION PERFORMANCE USING DIFFERENT EMBEDDING-CLASSIFIER COMBINATIONS

Classifier	Accuracy	F1_Score	Precision	Recall
DarijaBERT				
Logistic Regression	95.35%	95.35%	95.35%	95.35%
SVM	95.07%	95.07%	95.07%	95.07%
Random Forest	91.90%	91.90%	91.90%	91.90%
Mlp	96.00%	95.99%	96.00%	96.00%
FastText				
Logistic Regression	73.34%	73.13%	73.86%	73.34%
SVM	75.02%	74.68%	76.12%	75.02%
Random Forest	76.49%	76.43%	76.65%	76.49%
Mlp	76.32%	76.17%	76.80%	76.32%
Word2Vec				
Logistic Regression	73.42%	73.06%	74.40%	73.42%
SVM	73.34%	72.93%	74.50%	73.34%
Random Forest	74.52%	74.42%	74.76%	74.52%
Mlp	76.37%	76.19%	76.96%	76.37%

Table V summarizes the performance of the three embedding approaches across traditional classifiers. DarijaBERT achieves the highest scores consistently, particularly when used with logistic regression. FastText and Word2Vec also demonstrate reasonable performance, but fall short in capturing deep contextual features.

Although transformer-based models require more computational resources, their predictive performance justifies the trade-off in applications where accuracy is critical. Thus, we do not report execution time and memory footprint, as the focus of this study lies in evaluating classification quality.

VI. CHALLENGES AND FUTURE DIRECTIONS

A significant challenge in developing effective sentiment analysis models for Moroccan Darija lies in the scarcity of high-quality, annotated datasets. Unlike more widely used languages or even Modern Standard Arabic (MSA), Darija lacks the extensive labeled corpora necessary to train robust machine learning models. This data scarcity limits the ability of models to learn the full spectrum of linguistic features specific to Darija, including its unique syntax, vocabulary, and colloquial expressions. To achieve meaningful advancements in the field of sentiment analysis for this dialect, it is critical to focus on building larger, more diverse datasets that accurately represent the informal, regionally varied nature of Darija. This effort would not only provide more comprehensive coverage of the dialect's linguistic nuances but also offer richer training data for models that need to adapt to Darija's rapidly evolving vocabulary, particularly in digital communication.

Additionally, there is a pressing need for further research and development of advanced tokenization tools that can effectively handle the complexities of both Arabic script and Arabizi, the Latin-character transcription commonly used in informal online communication. Tokenization, which involves breaking down text into smaller units such as words or subwords, becomes particularly challenging in the context of Darija because of its non-standardized spelling and frequent code-switching between languages like Arabic, French, and Berber. These linguistic variations are further compounded when Darija is written in Arabizi, where the lack of formal rules for transliteration adds another layer of complexity.

The current limitations in tokenization tools often result in models struggling to parse Darija texts accurately, especially when they switch between scripts or contain unconventional spellings. As a result, improving tokenization methodologies for both Arabic script Darija and Arabizi is vital for enhancing model performance in sentiment analysis tasks. Moreover, the development of tokenization approaches tailored specifically to dialectal Arabic would greatly benefit not only Darija but also other Arabic dialects that share similar challenges. By investing in these advancements, the NLP community can build more accurate and effective models capable of navigating the linguistic diversity inherent in Moroccan Darija.

VII. CONCLUSION

This study highlights the effectiveness of transformer models in performing sentiment analysis on Moroccan Darija, underscoring the critical role that dialect-specific models like DarijaBERT play in accurately capturing the nuances of the language. By focusing on data that reflects the unique linguistic patterns of Darija, DarijaBERT outperformed more general models such as BERT Multilingual and AraBERT, both of which offered reasonable results but ultimately struggled to handle the complexities inherent to Darija. These complexities include its informal grammar, regional variations, and the influence of other languages like French, Berber, and Spanish, all of which contribute to the richness and variability of the dialect.

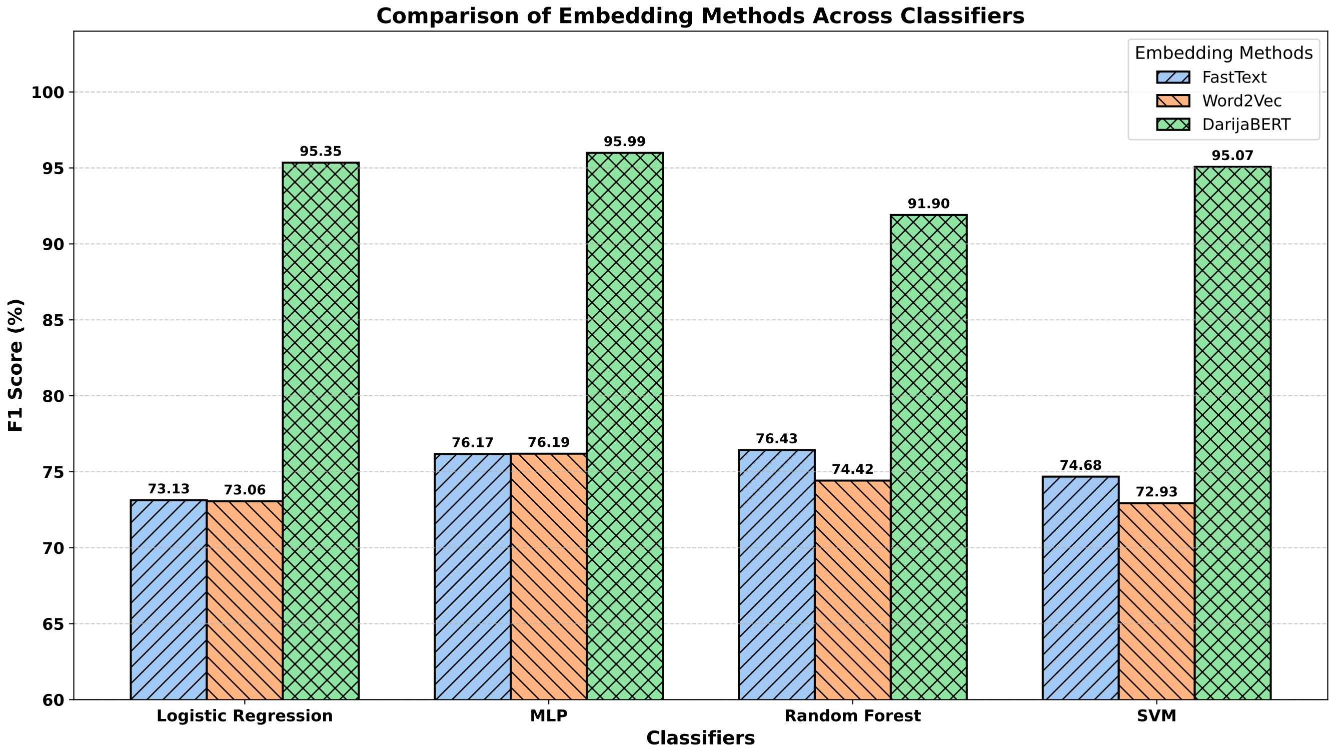

Fig. 7. Evaluating language models performance.

The superior performance of DarijaBERT highlights the importance of domain-specific pre-training, showing that models tailored to a particular dialect are far more effective at capturing its subtleties than models pre-trained on broader multilingual or MSA datasets. This success illustrates that, in the realm of sentiment analysis for low-resource dialects like Darija, generalized models are insufficient to fully understand the intricacies of the language. Instead, models that have been fine-tuned on dialect-specific data—such as DarijaBERT—can significantly improve performance by better recognizing colloquialisms, informal sentence structures, and culturally specific expressions that are unique to the dialect.

Moving forward, it is clear that future research should prioritize the expansion of resources for Moroccan Darija, including the development of larger, more comprehensive annotated datasets that cover the full range of linguistic variations within the dialect. Furthermore, there is a pressing need for the creation of more advanced tools and methodologies that can process dialectal Arabic more effectively. This includes the development of improved tokenization strategies for both Arabic script and Arabizi, as well as models capable of managing the fluidity and complexity of spoken dialects as they appear in digital text.

In conclusion, this study demonstrates that focusing on dialect-specific models can lead to substantial improvements in sentiment analysis tasks for languages like Moroccan Darija. Expanding the resources and tools available for these low-resource dialects will not only enhance model performance but also contribute to a deeper understanding of the linguistic diversity present within the Arabic-speaking world. Our findings demonstrate that contextual embeddings, particularly DarijaBERT, significantly outperform traditional approaches when used with neural classifiers such as MLP. The results highlight the importance of developing language-specific models for under-resourced dialects. Future research and development efforts should continue to explore innovative approaches to processing these dialects in order to meet the growing demand for more accurate and effective NLP solutions in diverse linguistic contexts.

Data Availability Statement

The data for arabic sentiment analysis MYC that support the findings of this study are openly available at https://github.com/MouadJb/MYC

Author Contributions

All authors were involved in the design and implementation of the research, as well as in analyzing the results and to the writing of the manuscript.

References

[1] Xingtong Ge, Xiaofang Jin, and Ying Xu, ”Research on sentiment analysis of multiple classifiers based on word2vec,” 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 2, pp. 230–234, 2018.

[2] Spraha Kumawat, Inna Yadav, Nisha Pahal, and Deepti Goel, ”Sentiment analysis using language models: A study,” 11th International Conference on Cloud Computing, Data Science and Engineering (Confluence), pp. 984–988, 2021.

[3] Shreyash Mishra, Siddhartha Choubey, Abha Choubey, N. Yogeesh, J. Durga Prasad Rao, and P. William, ”Data extraction approach using natural language processing for sentiment analysis,” International Conference on Automation, Computing and Renewable Systems (ICACRS), pp. 970–972, 2022.

[4] Maha Al-Ghalibi, Adil Al-Azzawi, and Kai Lawonn, ”NLP based sentiment analysis for Twitter's opinion mining and visualization,” 11th International Conference on Machine Vision, vol. 11041, pp618–626, 2019.

[5] Mourad Jabrane, Imad Hafidi, and Yassir Rochd, ”An Improved Active Machine Learning Query Strategy for Entity Matching Problem,” Advances in Machine Intelligence and Computer Science Applications, pp. 317–327, 2023.

[6] Mourad Jabrane, Hiba Tabbaa, Yassir Rochd, and Imad Hafidi, ”ERABQS: entity resolution based on active machine learning and balancing query strategy,” Journal of Intelligent Information Systems, vol. 62, no. 5, pp. 1347–1373, 2024.

[7] Mourad Jabrane, Abdelfattah Toulaoui, and Imad Hafidi, ”Enhancing Semantic Web Entity Matching Process Using Transformer Neural Networks and Pre-Trained Language Models,” Computing and Informatics, vol. 43, no. 6, pp. 1397–1415, 2024.

[8] Mourad Jabrane, Hiba Tabbaa, Aissam Hadri, and Imad Hafidi, ”Enhancing Entity Resolution with a Hybrid Active Machine Learning Framework: Strategies for Optimal Learning in Sparse Datasets,” Information Systems, vol. 125, pp. 102410, 2024.

[9] Ali Bou Nassif, Ashraf Elnagar, Omar Elgendy, and Yaman Afadar, ”Arabic fake news detection based on deep contextualized embedding models,” Neural Computing and Applications, vol. 34, no. 18, pp. 16019–16032, 2022.

[10] Khaled M. Fouad, Sahar F. Sabbeh, and Walaa Medhat, ”Arabic Fake News Detection Using Deep Learning,” Computers, Materials and Continua, vol. 71, no. 2, pp. 3647–3665, 2022.

[11] Mohammad Azzeh, Abdallah Qusef, and Omar Alabboushi, ”Arabic Fake News Detection in Social Media Context Using Word Embeddings and Pre-trained Transformers,” Arabian Journal for Science and Engineering, 2024.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, and Morgan Funtowicz, ”Transformers: State-of-the-art natural language processing,” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, 2020.

[13] Quoc Thai Nguyen, Thoai Linh Nguyen, Ngoc Hoang Luong, and Quoc Hung Ngo, ”Fine-tuning bert for sentiment analysis of vietnamese reviews,” 7th NAFOSTED conference on information and computer science (NICS), pp. 302–307, 2020.

[14] Hamdi Ali, Shaban Khaled, and Zainal Anazida, ”A Review on Challenging Issues in Arabic Sentiment Analysis,” Journal of Computer Science, vol. 12, no. 9, pp. 471–481, 2016.

[15] Alayba Abdulaziz M, Palade Vasile, England Matthew, and Iqbal Rahat, ”Improving sentiment analysis in Arabic using word representation,” 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), 2018.

[16] Ismail Rua, Omer Mawada, Tabir Mawada, Mahadi Noor, and Amin Izzeldein, ”Sentiment analysis for Arabic dialect using supervised learning,” IEEE International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018.

[17] Ratican Jay and Hutson James, ”Advancing Sentiment Analysis Through Emotionally-Agnostic Text Mining in Large Language Models (LLMS),” Journal of Biosensors and Bioelectronics Research, 2024.

[18] Zhang Tong, Gong Xinrong, and Chen CL Philip, ”BMT-Net: Broad Multitask Transformer Network for Sentiment Analysis,” IEEE Transactions on Cybernetics, vol. 52, no. 7, pp. 6232–6243, 2021.

[19] Delbrouck Jean-Benoit, Tits Noe, and Dupont Stephane, ”Modulated fusion using transformer for linguistic-acoustic emotion recognition,” arXiv preprint arXiv:2010.02057, 2020.

[20] Wu Zhen, Ying Chengcan, Dai Xinyu, Huang Shujian, and Chen Jiajun, ”Transformer-Based multi-aspect modeling for multi-aspect multi-sentiment analysis,” Natural Language Processing and Chinese Computing: 9th CCF International Conference, 2020.

[21] Horvat Marko, Gledec Gordan, and Leontic Fran, ”Hybrid Natural Language Processing Model for Sentiment Analysis during Natural Crisis,” Electronics, vol. 13, no. 10, pp1991, 2024.

[22] Qi Qingfu, Lin Liyuan, Zhang Rui, and Xue Chengrong, ”MEDT: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis,” IEEE Access, vol. 10, pp. 28750–28759, 2022.

[23] Xie Long-Fei and Zhang Xu-Yao, ”Gate-fusion transformer for multimodal sentiment analysis,” International Conference on Pattern Recognition and Artificial Intelligence, 2020.

[24] Nassif Ali Bou, Darya Abdollah Masoud, and Elnagar Ashraf, ”Empirical evaluation of shallow and deep learning classifiers for Arabic sentiment analysis,” Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 1, pp1–25, 2021.

[25] Yafoz Ayman and Mouhoub Malek, ”Analyzing machine learning algorithms for sentiments in arabic text,” IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020.

[26] Omara Eslam, Mosa Mervat, and Ismail Nabil, ”Applying recurrent networks for Arabic sentiment analysis,” Menoufia Journal of Electronic Engineering Research, vol. 31, no. 1, pp. 21–28, 2022.

[27] Matrane Yassir, Benabbou Faouzia, and Sael Nawal, ”Sentiment analysis through word embedding using AraBERT: Moroccan dialect use case,” International Conference on Digital Age and Technological Advances for Sustainable Development (ICDATA), 2021.

[28] Xu Lisha and Song Yuan, ”Comparison of text sentiment analysis based on traditional machine learning and deep learning methods,” 4th International Conference on Computer Engineering and Application (ICCEA), 2023.

[29] Thampy, S. Anjana, and Angelina Jeyaraj Jane Rubel, ”Deep Learning Architectures Based Sentiment Analysis Systematic Literature Review,” International Conference on Control, Communication and Computing (ICCC), 2023.

[30] Lal Chaman and Nasir Zafar, ”Comparative Analysis of Deep Learning Methods in the Realm of Sentiment Analysis,” International Multi-disciplinary Conference in Emerging Research Trends (IMCERT), vol. 1, 2023.

[31] Singh Jitendra and Sharma Geeta, ”Sentiment Analysis Study of Human Thoughts using Machine Learning Techniques,” International Conference on Disruptive Technologies (ICDT), 2023.

[32] Qamar Mohammad, Rao Hamnah, Farooq Sheikh Afaan, and Bhuyan Ajatray Swagat, ”Sentiment analysis using deep learning: A domain independent approach,” Second International Conference on Electronics and Renewable Systems (ICEARS), 2023.

[33] Kumar Narinder, Kaur Kiranpreet, Saini Rupinder, Singla Sanjay, and Shilpa, ”Evaluation of Sentiment using Deep Learning and Machine Learning using Word Integration Techniques,” First International Conference on Technological Innovations and Advance Computing (TIACOMP), 2024.

[34] Moudhich Ihab and Fennan Abdelhadi, ”Graph embedding approach to analyze sentiments on cryptocurrency,” International Journal of Electrical and Computer Engineering, pp. 690–697, 2024.

[35] Ambinintsoa Lorenzo Mamelona, Kyeremeh Bright Bediako, and Osibo Benjamin Kwapong, ”Unveiling Sentiments: A Comparative Study of Machine Learning-based Word Representations for Enhanced Sentiment Analysis,” 2023.

[36] Rong Huang, Qianyi Chen, Jun Tang, and Jianjie Song, ”The Influence of Word Embeddings on the Performance of Sentiment Classification,” International Journal of Computer and Information Technology, vol. 1, no. 4, pp1–1, 2023.

[37] Peter Atandoh, Paul Atandoh Hakeem, Edward Mensah Acheampong, Daniel Addo, Michael Appiah-Twum, Daniel Adu-Gyamfi, Richard Safo Blankson, and Tohye Tewodros Giza Wi, ”Enhanced Word Embedding with CNN using Word Order Information For Sentiment Analysis,” Proceedings of ICCWAMTIP, 2023.

[38] Lavanya B.N, Anitha Rathnam K.V, K. K., Abhishek Appaji, P. D. Shenoy, and Venugopal K. R, ”Sentiment Analysis of Textual Data using Word Embedding and Deep Learning Approaches,” Proceedings of NKCON, 2023.

[39] Surapaneni Pavan Nikhith, Popuri Varun Kumar, Aira Udaybhasker, T. Raghunadha Reddy, P. Reddy, and Tripty Singh, ”Sentiment Analysis of Airline Tweets Using Word Embeddings and Deep Learning Techniques,” Proceedings of ICCCNT, pp. 1–7, 2024.

[40] Gavali Prashantkumar M and Shiragave Suresh K, ”Text Representation for Sentiment Analysis: From Static to Dynamic,” Proceedings of ICSMDI, 2023.

[41] Hani Almaqtari, Feng Zeng, and Ammar Mohammed, ”Enhancing Arabic Sentiment Analysis of Consumer Reviews: Machine Learning and Deep Learning Methods Based on NLP,” Algorithms, 2024.

[42] Martín Serrano, Hager Saleh, Ali Abdullah Hamzah, Nora El-Rashidy, Abdullah Alharb, Ahmed Elaraby, and Sherif Mostafa, ”ArabBert-LSTM: Improving Arabic Sentiment Analysis Based on Transformer Model and Long Short-Term Memory,” Frontiers in Artificial Intelligence, vol. 7, 2024.

[43] Ali Rachidi, Ali Ouacha, and Mohamed El Ghmary, ”Sentiment Analysis by Deep Learning Techniques,” Lecture Notes in Computer Science, 2024.

[44] Giuseppe Varone, Rami K. Ahmed, Mandar Gogate, Kia Dashtipour, Hani Almoamari, Mohammed Ahmed El-Affendi, Bassam Naji Al-Tamimi, Faisal Albalwy, and Amir Hussain, ”Arabic Sentiment Analysis Based on Word Embeddings and Deep Learning,” Computers, vol. 12, no. 6, pp. 126–126, 2023.

[45] Sahar F. Sabbeh and Heba Fasihuddin, ”A Comparative Analysis of Word Embedding and Deep Learning for Arabic Sentiment Classification,” Electronics, vol. 12, no. 6, pp. 1425–1425, 2023.

[46] Pan Xie, Hengnian Gu, and Dongdai Zhou, ”Modeling Sentiment Analysis for Educational Texts by Combining BERT and FastText,” Proceedings of CSTE, 2024.

[47] Nouri Hicham, Habbat Nassera, and Sabrina Karim, ”Enhancing Arabic E-Commerce Review Sentiment Analysis Using a Hybrid Deep Learning Model and FastText Word Embedding,” EAI Endorsed Transactions on Internet of Things, 2023.

[48] Ibrahim Kaibi, El Habib Nfaoui, and Hassan Satori, ”Sentiment Analysis Approach Based on Combination of Word Embedding Techniques,” Proceedings of the Conference on Advances in Intelligent Systems and Computing, 2020.

[49] M. Raihan, Erwin Budi Setiawan, and Jurnal, ”Aspect Based Sentiment Analysis with FastText Feature Expansion and Support Vector Machine Method on Twitter,” Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 6, no. 4, pp. 591–598, 2022.

[50] Ibrahim Kaibi, El Habib Nfaoui, and Hassan Satori, ”A Comparative Evaluation of Word Embeddings Techniques for Twitter Sentiment Analysis,” Proceedings of WITS, 2019.

[51] Ruoyu Liu, ”Exploring the Impact of Word2Vec Embeddings Across Neural Network Architectures for Sentiment Analysis,” Applied and Computational Engineering, vol. 97, no. 1, pp. 93–98, 2024.

[52] Muhammad Dimas Rifki Irianto, Mahendra Dwifebri Purbolaksono, and Bunyamin Bunyamin, ”Sentiment Analysis of Livin’ by Mandiri Application Reviews Using Word2Vec Feature Extraction and KNN Method,” Proceedings of the International Conference on Data Science and Applications (ICoDSA), pp. 236–241, 2024.

[53] Alaaddin Goktug Ayar, Sercan Aygun, M Hassan Najafi, and Martin Margala, ”Word2HyperVec: From word embeddings to hypervectors for Hyperdimensional computing,” Proceedings of the Great Lakes Symposium on VLSI, 2024.

[54] Albert Vinluan, ”Research on Emotional Analysis of Online Book Reviews Based on Word2Vec Method,” Frontiers In Business, Economics and Management, 2023.

[55] Yahui Wang, Xiaoqing Cheng, and Xuelei Meng, ”Sentiment analysis with an integrated model of BERT and bi-LSTM based on multi-head attention mechanism,” International Journal of Computer Science, vol. 50, no. 1, pp. 255–262, 2023.

[56] Venkatesh, Siddhanth U Hegde, Satish B Basapur, and Nagaraju Y, ”DistilBERT-CNN-LSTM Model with GloVe for Sentiment Analysis on Football Specific Tweets,” IAENG International Journal of Computer Science, vol. 49, no. 2, pp. 420-432, 2022.

[57] Carlos Henríquez, Freddy Briceño, and Dixon Salcedo, ”Unsupervised model for aspect-based sentiment analysis in Spanish,” IAENG International Journal of Computer Science, vol. 46, no. 3, pp. 430–438, 2019.

[58] Shihab Elbagir and Jing Yang, ”Sentiment analysis on Twitter with Python’s natural language toolkit and VADER sentiment analyzer,” IAENG Transactions on Engineering Sciences: Special Issue for the International Association of Engineers Conferences, 2020.

[59] Cong Cuong Le, PWC Prasad, Abeer Alsadoon, Linh Pham, and A Elchouemi, ”Text classification: Naïve bayes classifier with sentiment Lexicon,” International Journal of Computer Science, vol. 46, no. 2, pp. 141–148, 2019.

[60] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao, ”Learning deep transformer models for machine translation,” arXiv preprint arXiv:1906.01787, 2019.

[61] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou, ”Achieving Human Parity on Automatic Chinese to English News Translation,” arXiv, 2018.

[62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv, 2018.

[63] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, ”SQuAD: 100,000+ Questions for Machine Comprehension of Text,” arXiv, 2016.

[64] Guillaume Lample and Alexis Conneau, ”Cross-lingual Language Model Pretraining,” arXiv, 2019.

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, ”Attention Is All You Need,” arXiv preprint arXiv:1706.03762, vol. 10, 2017.

[66] Alrefai Mo’ath, Faris Hossam, and Aljarah Ibrahim, ”Sentiment analysis for Arabic language: A brief survey of approaches and techniques,” arXiv preprint arXiv:1809.02782, 2018.

[67] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao, ”Multi-Task Deep Neural Networks for Natural Language Understanding,” arXiv, 2019.

[68] Delbrouck Jean-Benoit, Tits Noe, and Dupont Stephane, ”Modulated fusion using transformer for linguistic-acoustic emotion recognition,” arXiv preprint arXiv:2010.02057, 2020.