A Study of Sentiment Analysis Applied to Moroccan Dialect

The National School of Applied Sciences in Al Hoceima,
Abdelmalek Essaadi University

Author: EL HADRATI Othman
Supervision: KHAMJANE Aziz

othman.elhadrati@etu.uae.ac.ma

Abstract

Social media has become a dominant form of communication, with users frequently expressing opinions on various topics. **Sentiment analysis** of social media content, particularly tweets or comments, offers a valuable method for understanding public opinion on contents, events, and other subjects. **By analyzing the emotions** transmitted in these online expressions, governments can identify societal concerns, such as public reactions to policies, and individuals can enhance their decision-making, such as tailoring marketing strategies or improving customer experiences. This work focuses on enhancing sentiment classification specifically within the Moroccan context.

Keywords— sentiment analysis, Moroccan Arabic dialect, social media, machine learning

1 Introduction

Natural Language Processing (NLP) is a crucial branch of **artificial intelligence** focused on enabling computers to understand and process human language, both written and spoken. This field tackles the complexities inherent in language, including variations in grammar, syntax, and semantics. A significant challenge within NLP is the processing of dialects, which often differ substantially from standard language. This research paper addresses this challenge by focusing on the **Moroccan dialect, Darija**. **Darija**, with its unique linguistic characteristics and growing presence in online communication, presents a compelling case study for **NLP research**.

One important application of **Natural Language Processing (NLP)** is **sentiment analysis (SA)**, which aims to identify and categorize the **sentiment polarity** expressed in text. **SA** plays a vital role in understanding public opinion, customer feedback, and social trends. By analyzing text, **SA** can determine whether the sentiment expressed is **positive** or **negative**. This research leverages the power of **NLP** and **SA** to address the specific linguistic challenges posed by **Darija**, aiming to develop improved **sentiment analysis models** for this dialect. This is particularly important given the increasing use of **Darija** in social media and online platforms, where understanding public sentiment can provide valuable insights.

2 Related Works

Several research efforts have explored various aspects of Natural Language Processing (NLP) for Moroccan Darija. While sentiment analysis has been a prominent area of focus, other NLP tasks have also received attention.

In the study by **El Wardani Dadi, Sara Ouahab** and **El Ouahabi Safâa** [16], the authors constructed a large, multi-domain dataset of Moroccan Darija text collected from various social media platforms. This dataset facilitated their exploration of different machine learning approaches, including BERT, demonstrating the potential of these techniques for Darija sentiment analysis.

More recent research has leveraged advancements in machine learning. **Motassim Hamza**[9] introduced a machine learning and deep learning approach combining Modern Standard Arabic (MSA) and Darija for sentiment analysis, demonstrating improved performance. Notably, their work addressed the code-switching phenomenon, a crucial aspect of Moroccan social media text that impacts various NLP tasks beyond just sentiment analysis.

Mohamed Amine Ouassil and **Rabia Rachidi** [13] addressed sentiment analysis of Twitter comments written in both Modern Standard Arabic and Moroccan Dialectal Arabic. They employed a machine learning approach, combining various feature extraction techniques (n-grams, BOW/TF-IDF, word embeddings) with different classifiers (Naive Bayes, Random Forests, SVM, Logistic Regression, and LSTM). Their experiments demonstrated the potential of this approach, with the SVM model achieving an accuracy close to 70% .

This research specifically contributes to the literature by focusing on sentiment analysis for Moroccan

Darija. By addressing the unique challenges of non-standardized dialects, this work aims to advance the broader field of sentiment analysis and enhance its real-world applicability.

3 Model Architecture

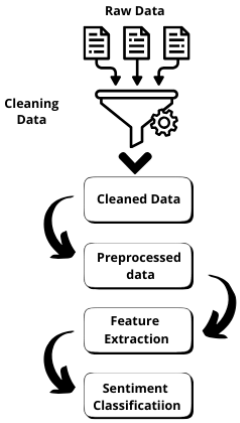

Figure 1: Proposed Architecture for Darija Sentiment Analysis: Pipeline showing data preprocessing, feature extraction, and classification stages for analyzing Moroccan dialect text

4 Methodology

This chapter details the methodology employed for classifying sentence sentiment. It describes the text models, dataset, and classifiers used, along with the preprocessing and normalization techniques designed to handle the informal nature of dialectal Arabic, particularly Moroccan Darija. The evaluation metrics used to assess sentiment classification performance are also presented. The methodology can be summarized as follows: a description of the dataset used in this work; the application of various preprocessing steps, including normalization, noise removal, and conversion of emoticons to text, to enhance classification performance; an explanation of the techniques used for text representation; and finally, the classification of Moroccan dialect text using machine learning, deep learning, and fine-tuned pretrained model classifiers, followed by a comparison of the results.
To demonstrate the functionality of the sentiment analysis model, we deployed it in a web application, which can be accessed at [15]. The app allows users to input text or Youtube video Link and receive real-time sentiment predictions based on the model.

4.1 Data Collection

The data collection process involved multiple sources to ensure diversity in the dataset:

Existing Datasets: Five different pre-existing datasets were leveraged:

- Dataset 1 [8] **MAC (Moroccan Arabic Corpus)**: – A free and large corpus consisting of 18,000 manually labeled tweets, MAC is the first and largest Moroccan Arabic corpus for sentiment analysis, praised for its size, and accessibility to the research community. It provides a benchmark for future work, including polarity classification and language identification.
- Dataset 2 [5] **OMCD (Offensive Moroccan Comments Dataset)**: – This dataset focuses on detecting offensive content, such as verbal attacks and hate speech, in Moroccan Arabic on social media. It is the first of its kind for the Moroccan dialect, offering annotated data for detecting offensive language in Dialectal Arabic (DA).
- Dataset 3 [10] **Moroccan Darija Offensive Language Detection Dataset**: – This dataset contains 20,402 Moroccan Darija sentences, labeled as either offensive or non-offensive, and gathered from Twitter and YouTube comments.
- Dataset 4 [4] **ElecMorocco2016**: – A sentiment analysis dataset containing 10,254 Arabic Facebook comments related to the 2016 Moroccan elections. The comments are written in both Standard Arabic and Moroccan Dialect (Darija), offering insights into political sentiments during the election period.
- Dataset 5 [2] **Covid-19 Sentiment Analysis Dataset**: – A dataset containing Facebook comments during the Covid-19 pandemic, focused on sentiment analysis and public opinion during this global crisis. The dataset includes comments in both Standard Arabic and Moroccan Dialect (Darija).

YouTube API Data Collection: The dataset was enriched by:
- Extracting comments from popular Moroccan YouTube channels.
- Focusing on videos with high engagement rates.
- Collecting comments from diverse topics (reactions, programs, lifestyle).

Data Processing Steps for the collected comments:
- Preprocessing and cleaning the collected text.
- Removing duplicates and irrelevant content.
- Using K-means clustering for sentiment annotation, where the DarijaBERT-Arabizi pre-trained model [6] was used to obtain sentence embeddings. Specifically, the embeddings were extracted from the last hidden layer of the pre-trained model [11], capturing contextual representations of the text. These embeddings were then utilized to perform clustering and identify similarities between the collected text.

- Validation of clustering results through human review.

The Production dataset used for training and

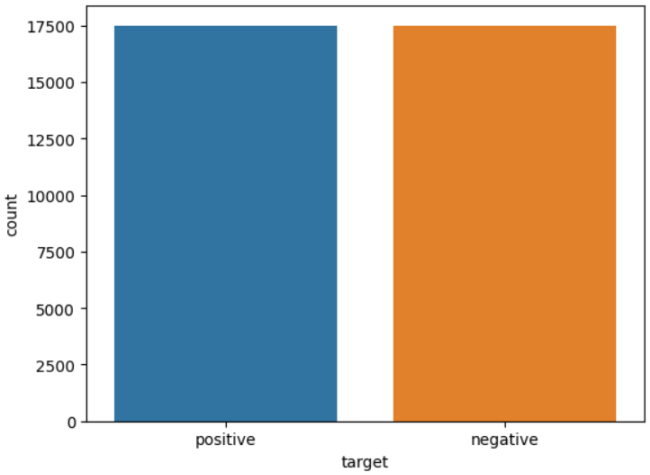

Figure 2: Balanced target distribution

evaluation is balanced, ensuring a fair representation of both positive and negative sentiments.

4.2 Data Cleaning and Preprocessing

Cleaning and preprocessing techniques are fundamental in Sentiment Analysis (SA) for Arabic dialectal text, particularly for the Moroccan Darija dialect. This necessity arises from the unique challenges presented by social media text, which is characterized by its informal nature and lack of standardization. The implementation addresses these challenges through a comprehensive preprocessing pipeline specifically designed for Darija text, operating at multiple levels of text normalization.

The preprocessing approach is distinguished by its specialized handling of Darija's dual-script nature (Latin and Arabic) and its complex morphological structure. The preprocessing pipeline consists of the following steps:

- **Class Adaptation:** Ensuring that data across different datasets is consistent by aligning classes and removing any empty fields. This ensures that the dataset is clean and usable for further analysis.
- **Stop Words Removal:** Removal of stop words in both Arabic script and Arabizi (Latin script) representations of Darija words. This dual-script approach ensures comprehensive removal of non-informative words regardless of their writing system. The stop words lists are sourced from various reputable lexicons, including:
  - Arabic Stop Words [1] and [3] .
  - Moroccan stop words [18] and [7] .
  - DODa stop words (Arabizi , arabic)[17] .

4.2.1 Data Cleaning

Cleaning is focused on eliminating irrelevant or noisy elements from the dataset. This step ensures that the data is free from non-informative parts, which could negatively affect the model's performance. The cleaning steps include:

- **Noise Removal:** Removal of special characters, punctuation marks, numbers in both Arabic and Latin scripts, redundant or spamming alphabets, and other unwanted symbols that do not contribute to the sentiment analysis task.
- **Context-Aware Cleaning:** Handling modern social media elements, including:
  - Emojis translation to text using a dedicated library.
  - Removal of URLs, Emails, underscores, special tags (e.g., &, ..), and HTML tags.
  - Removal of YouTube timestamps.

4.2.2 Text Preprocessing

Preprocessing techniques prepare the text for effective analysis by transforming it into a standard format. This step includes:

- **Custom Latin-to-Arabic Mapping:** A character mapping system that handles various romanized representations of Arabic phonemes for exemple :

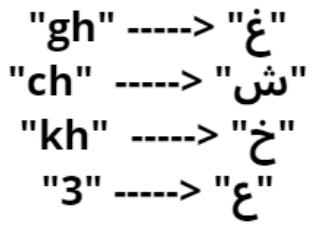

Figure 3: Darija transliteration mapping

- **Normalization:**
  - Removal of diacritical marks (tashkeel).
  - Elimination of elongation characters (tatweel).
  - Standardization of various forms of hamza and alef.
  - Removal of definite article "AL" prefix.
- **Stemming:** Application of Arabic-specific stemming techniques using the Tashaphyne light stemmer and custom suffixes and prefixes scraped from [12] specific to Moroccan Darija to reduce morphological variants to their base forms.

UnNormalized Arabic Letters	Normalized Letters
Hamza(ء,ئ)	ء
Alif(ا,إ,آ,أ)	ا
Alif Lam(ال, أل)	ال
Hae(ة)	ه
Dad(ض,ظ)	ض

Figure 4: Normalization of some arabic letters

This comprehensive approach ensures that the text maintains its semantic integrity while being standardized for effective sentiment analysis, resulting in a balanced dataset with over 17,500 positive sentences and a similar number of negative sentences.

4.3 Feature Extraction

Feature extraction is a critical step in transforming textual data into numerical representations for analysis. Several approaches are utilized in this context, including Bag of Words, Binary Bag of Words, and Term Frequency-Inverse Document Frequency (TF-IDF).

4.3.1 Bag of Words

The Bag of Words (BoW) model generates a vector representation of text by counting the frequency of each word's occurrence. BoW is frequently employed in text classification and document recognition tasks due to its simplicity and effectiveness. However, this approach does not account for the importance of rare words or their distribution across the corpus.

4.3.2 Binary Bag of Words

The Binary Bag of Words (Binary BoW) is a variation of the BoW model, where the presence or absence of a word in a document is recorded as a binary value (1 if the word appears, 0 if it does not). This method simplifies the representation by focusing solely on the occurrence of words, ignoring their frequencies.

4.3.3 TF-IDF

To address the limitations of BoW, Term Frequency-Inverse Document Frequency (TF-IDF) is employed. This method assigns weights to words based on their relevance in a document and their distribution across the entire corpus. Words that appear frequently in a document but rarely in the corpus are given higher weights, enhancing their discriminative power.

Task	Result
Original Text	احسن حكومة في المغرب ساصوت على العدالة والتنمية اناس شرفاء لم يسرقوا بلدنا النصر حليفهم ان شاء الله
Cleaned text	['حليفهم', 'حكومه', 'بلدنا', 'نسر', 'ان' 'اناس', 'مغرب', 'له', 'ساصوت', 'في', 'احسن', 'لم', 'والتنمية', 'يسرقوا' 'شرفا', 'شاا', 'علي', 'عداله']
Uni-gram n=1	['احسن', 'حليفهم', 'ساصوت', 'شرفاا', 'عداله', 'لم', 'نسر' 'والتنمية', 'يسرقوا']
Uni-gram n=2	['احسن', 'احسن لم', 'حليفهم', 'حليفهم نسر' 'ساصوت', 'ساصوت احسن','شرفاا', 'شرفاا عداله', 'عداله', 'لم' لم والتنمية', 'نسر','نسر ساصوت', 'والتنمية', 'والتنمية يسرقوا', 'يسرقوا','يسرقوا شرفاا]
Uni-gram n=3	['احسن', 'احسن لم', 'احسن لم والتنمية', 'حليفهم', 'حليفهم نسر' 'حليفهم نسر ساصوت', 'ساصوت', 'ساصوت احسن', 'ساصوت احسن لم' 'شرفاا', 'شرفا عداله', 'عداله', 'لم', 'لم والتنمية' 'لم والتنمية يسرقوا', 'نسر', 'نسر ساصوت', 'نسر ساصوت احسن' 'والتنمية', 'والتنمية يسرقوا', 'والتنمية يسرقوا شرفاا', 'يسرقوا' 'يسرقوا شرفاا', 'يسرقوا شرفا عداله']

Figure 5: Features Extraction

The formulas for computing TF and IDF are as follows:

Term Frequency (TF):

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

Inverse Document Frequency (IDF):

$$\mathrm{idf}_i = \log \frac{N}{\mathrm{df}_i} \tag{2}$$

TF-IDF Weight:

$$w_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i \tag{3}$$

Where:

- $n_{i,j}$ is the number of occurrences of term i in document j.
- df_i is the number of documents containing the term i.
- N is the total number of documents in the corpus.

TF measures the frequency of a term in a document, while IDF reduces the weight of common terms across the corpus. The resulting TF-IDF matrix has dimensions (documents × vocabulary), capturing the importance of terms across the entire dataset.

This approach allows for effective handling of common and rare terms, ensuring robust text analysis.

5 Classification Models

Several classification algorithms were tested and tuned to perform sentiment analysis. Each model brought valuable insights, and although multiple approaches showed promising results, the Support Vector Machine (SVM) with TF-IDF and bigrams ultimately delivered the best performance. All models were carefully tuned using hyperparameter optimization techniques to ensure their best possible performance. Below, we describe the models tested and explain why the SVM model was selected.

5.1 Machine Learning Models

5.1.1 Logistic Regression

Logistic Regression is a linear model that predicts the probability of a class using a sigmoid function. It's often used for binary classification tasks due to its simplicity and interpretability. Although it performed reasonably well as a baseline model, its performance was limited by its inability to capture complex dependencies between words in the text.

5.1.2 Multinomial Naive Bayes

The Multinomial Naive Bayes classifier is particularly suited for text classification, where features represent word frequencies or counts. It is a probabilistic model based on Bayes' Theorem, assuming independence between features. While it performed decently, its assumption of feature independence limited its ability to model more complex relationships in the text.

5.1.3 Decision Tree

A Decision Tree classifier uses a tree-like structure to make predictions by recursively splitting the data based on feature values. It is easy to interpret and visualize but can be prone to overfitting, especially with high-dimensional data like text.

5.1.4 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy. It generally performs well with high-dimensional data and helps to reduce overfitting. While Random Forest performed adequately, it did not outperform the selected SVM model in terms of accuracy and efficiency for this specific problem.

5.1.5 XGBoost

XGBoost is a gradient boosting algorithm that iteratively corrects the errors made by previous models, often leading to very high accuracy. However, it requires significant computational resources and hyperparameter tuning, making it less efficient compared to the selected SVM model.

5.2 Deep Learning Models

5.2.1 Convolutional Neural Networks (CNNs)

CNN is effective at capturing local patterns and dependencies in the text, especially in shorter sequences. However, CNNs were not able to outperform traditional machine learning models for this task, likely due to the relatively small size and structure of the dataset.

5.2.2 Recurrent Neural Networks (RNNs)

RNN is designed for sequential data, maintaining hidden states that capture contextual dependencies between words. While they were able to capture more complex relationships in the text, they required more computational power and did not significantly improve upon the results of simpler models for this specific sentiment analysis task.

5.3 Pre-trained Large Language Models

In recent years, pre-trained large language models (LLMs) have demonstrated significant potential in various natural language processing (NLP) tasks, including sentiment analysis and text classification. These models, particularly those based on the transformer architecture, have been fine-tuned for specific languages and dialects to enhance their performance on specialized tasks. For Moroccan Darija, which presents unique challenges such as informal language, code-switching, and a lack of standardization, several pre-trained models have been explored to address these complexities.

While models like BERT have shown success across multiple languages, applying them to Darija requires careful adaptation to its unique structure. From my experience, fine-tuning these models for Darija has proven to be particularly challenging due to the informal and non-standardized nature of the dialect. Additionally, the need for substantial computational resources and high-quality labeled data has been a significant barrier. The process of fine-tuning for Darija is further complicated by code-switching between Arabic, French, and other languages, which is common in social media texts. Pre-trained models specifically designed for Moroccan Arabic or Arabizi text offer a more tailored approach to capture these linguistic features. Notable examples include DarijaBERT-Arabizi [6], DarijaBERT [6], and MorrBERT [14], each showing varying degrees of success in processing Moroccan Darija text. However, despite promising results, fine-tuning these models has not consistently led to superior performance compared to traditional machine learning models.

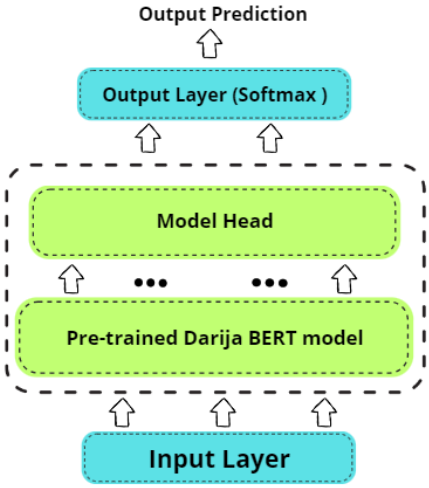

Figure 6: BERT Architecture

5.4 Selected Model: Support Vector Machine (SVM) with TF-IDF Vectorizer

The Support Vector Machine (SVM) model with TF-IDF and bigrams emerged as the best-performing model for this sentiment analysis task. This model was chosen due to its excellent balance between accuracy and computational efficiency. The following parameters were used for the SVM model after tuning the hyperparameters using GridSearch:

- $C = 1.0$: The regularization parameter, which controls the trade-off between achieving a low error on the training data and minimizing the margin.
- `kernel='rbf'`: The radial basis function kernel, which allows the model to handle non-linear decision boundaries effectively.
- `gamma='scale'`: A kernel coefficient value that adapts to the data, ensuring optimal performance.

5.4.1 Formal Description

The SVM algorithm aims to find the optimal hyperplane that maximizes the margin between the two classes. The optimization problem is formulated as:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

subject to the constraints:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, n$$

where w is the weight vector, b is the bias, ξ_i are the slack variables, and C is the regularization parameter.

The hinge loss function, which SVM uses during training, is defined as:

$$L(y_i, f(x_i)) = \max(0, 1 - y_i f(x_i)),$$

where $f(x_i) = w^T x_i + b$ is the decision function. This loss penalizes misclassified points and those within the margin, ensuring that the model not only classifies points correctly but also maintains a margin around the hyperplane.

The TF-IDF vectorizer was used to convert text data into numerical features, and bigrams were incorporated to capture contextual relationships between adjacent words. The use of bigrams allowed the model to better capture the nuances of sentiment in text, especially for informal dialects like Darija. TF-IDF weighted the importance of words based on their frequency and relevance, improving the model's ability to distinguish between relevant features and noise.

5.4.2 Why SVM with TF-IDF and Bigrams?

SVM with TF-IDF and bigrams provided several advantages over other models:

- **Efficiency:** SVM is computationally efficient compared to deep learning models, particularly when the dataset is not very large.
- **Effectiveness with Text Data:** The combination of TF-IDF and bigrams allowed the model to capture important word pair relationships that are crucial for understanding sentiment in informal text.
- **Interpretability:** SVM offers interpretability, making it easier to understand the decisions made by the model.
- **Scalability:** SVM with TF-IDF and bigrams can scale well to larger datasets with minimal loss in performance.

While other models were tested, SVM with TF-IDF and bigrams outperformed them in terms of both accuracy and efficiency, making it the most suitable choice for this sentiment analysis task.

6 Evaluation

The evaluation of the model was conducted using standard classification metrics to assess its performance comprehensively. The metrics used are as follows:

- **Accuracy:** Measures the proportion of correct predictions (both true positives and true negatives) out of the total number of predictions. In this case, the accuracy is 0.77, indicating that 77% of predictions were correct.

- **Precision:** Precision is the proportion of positive predictions that are actually correct. It is particularly important in scenarios where the cost of false positives is high.

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP (True Positives) represents the correctly identified positive cases, and FP (False Positives) represents the cases incorrectly identified as positive.

- **Recall:** Also known as sensitivity, recall measures the proportion of actual positives that are correctly identified. It is crucial when the cost of false negatives is high.

$$\text{Recall} = \frac{TP}{TP + FN}$$

where FN (False Negatives) represents the positive cases incorrectly identified as negative.

- **F1-Score:** The F1-score is the harmonic mean of precision and recall, providing a balance between them. It is particularly useful when there is an imbalance between classes.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The results for each class are as follows:

- **Positive class:** Precision = 0.76, Recall = 0.79, F1-Score = 0.78, Support = 3443.

- **Negative class:** Precision = 0.77, Recall = 0.74, F1-Score = 0.76, Support = 3322.

Confusion Matrix: The confusion matrix provides a detailed breakdown of the model's performance. It is structured as follows:

- **True Positives (TP):** Correctly identified positive cases (2728).

- **False Positives (FP):** Cases incorrectly identified as positive (715).

- **True Negatives (TN):** Correctly identified negative cases (2458).

- **False Negatives (FN):** Cases incorrectly identified as negative (864).

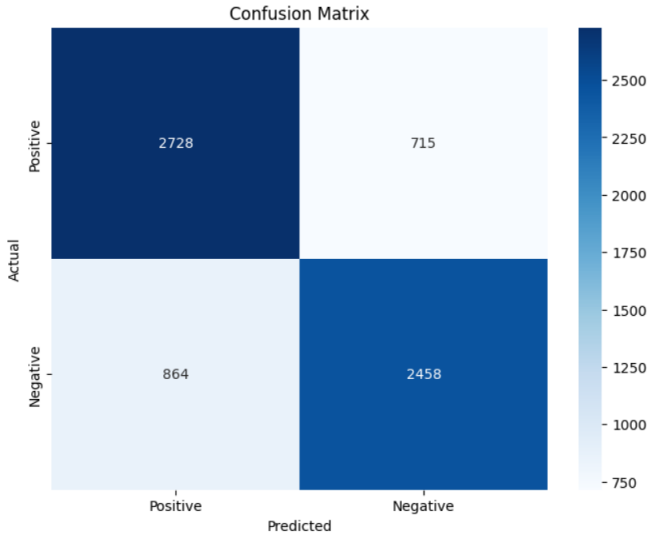

Figure 7: Confusion Matrix for the SVM Model

Class	Precision	Recall	F1-Score	Support
Positive	0.76	0.79	0.78	3443
Negative	0.78	0.75	0.76	3443

Figure 8: Classification Performance Evaluation Metrics

The confusion matrix and table of Performance (Figure 7, 8) highlights how well the model performed on each class and provides insights into areas where the model can be improved.

Figure 9: Positive class word occurrence

7 Model Performance and Benchmark

7.1 Testing of Models Results

The performance of various ML classifiers using different feature extraction techniques is summarized in Figure 10. Similarly, the accuracy of deep learning models and fine-tuned pre-trained models is presented in Figure 11.

Categorie	Classifier	TF-IDF			BoW			BoW (Binary)		
		Unigram	1g+2g	1g+2g+3g	Unigram	1g+2g	1g+2g+3g	Unigram	1g+2g	1g+2g+3g
ML	LR	0.75	0.75	**0.75**	0.74	0.74	0.74	**0.75**	**0.75**	**0.75**
	DT	0.71	0.70	0.71	0.71	0.70	0.70	0.70	0.71	0.70
	SVM	**0.77**	**0.77**	0.74	**0.75**	**0.75**	0.74	**0.75**	**0.75**	0.74
	RF	0.75	0.74	0.74	**0.75**	0.73	0.74	**0.75**	0.73	0.74
	XGBoost	0.71	0.71	0.71	0.72	0.72	0.72	0.72	0.72	0.72
	NB	0.75	0.75	**0.75**	**0.75**	0.73	**0.75**	**0.75**	**0.75**	**0.75**

Figure 10: Accuracy of ML classifiers with different feature extraction methods (TF-IDF, BoW, and BoW Binary).

Categorie	Classifier	Word Embedding
DL	CNN	0.756
	RNN	0.757
Pre-Trained models	DarijaBERT -Arabizi	0.515
	DarijaBERT	0.5010
	MorrBERT	0.7289

Figure 11: Accuracy comparison of DL models and fine-tuned pre-trained models.

7.2 Benchmark Comparison

The table below compares the performance metrics of the models based on accuracy, F1-score, precision, and recall.

Model	Precision	Recall	F1-Score	Accuracy
SVM	0.77	0.77	0.77	0.77
Hamza Model	0.74	0.74	0.74	0.74

Figure 12: Performance Comparison of Models

8 Limitations of the Approach

Despite the growing presence of Darija in written form across various platforms, processing it poses several unique challenges. These difficulties arise from the specific characteristics of the language, which complicate its handling in natural language processing (NLP) tasks. Some of the key challenges include:

- **Lack of Standardization:** Darija lacks a standardized orthography, leading to variations in spelling and writing conventions. For example:
  - The word "Good" can be written as "mzian," "mezyan," or "meziane."
  - "What" can be written as "chno," "chnou," "chnowa," or "cheni."
- **Code-Mixing:** Darija frequently mixes with French, English, and Modern Standard Arabic (MSA), creating code-switched text that complicates language processing. For example:
  - "khoya had weekend ghadi nmchiw ndiro shopping" (My brother, we'll go shopping this weekend)
  - "C'est bon, fhemtek" (It's good, I understand)
- **Informal Nature:** The informality of Darija, especially on social media, complicates the application of traditional NLP techniques. For example:
  - Using numbers as letters: "3ayane" instead of "aayane" (tired)
  - Elongating letters for emphasis: "mziaaaaaaan" instead of "mzian" (good)
  - Using abbreviations: "wlh" instead of "wallah" (I swear)

These challenges are not limited to sentiment analysis but affect a wide range of NLP applications, including machine translation, information retrieval, and text summarization.

9 Conclusion

This study investigated sentiment analysis for Moroccan Darija, addressing the unique linguistic challenges of this non-standardized Arabic dialect. Contributing to the broader field of NLP in informal linguistic contexts, we demonstrated the effectiveness of both traditional and deep learning approaches. Using a semi-supervised labeling approach leveraging k-means clustering on embedding vectors for

initial labeling followed by meticulous manual correction to have a valuable labeled resource. Experiments showed that traditional machine learning models (Multinomial Naive Bayes, Logistic Regression, XGBoost, SVM) and deep learning models (CNN and RNN) achieved promising results after hyperparameter optimization. However, fine-tuned BERT-based models (like DarijaBERT) did not yield the expected performance gains, suggesting the need for further research into optimal fine-tuning, architectural modifications, alternative pre-training, larger datasets, or alternative architectures better suited to Darija. We acknowledge limitations including potential data biases and computational demands. Despite these, this research offers valuable insights into Darija sentiment analysis and underscores the importance of addressing linguistic diversity in real-world NLP applications.

Looking ahead, we plan to further enhance our sentiment analysis model by integrating more advanced deep learning techniques, such as LSTM, and experimenting with fine-tuning strategies to improve performance. We will also explore alternative feature extraction methods that could lead to better accuracy. Additionally, we aim to broaden our dataset by including a wider variety of Moroccan Arabic sources and incorporating additional topics to enhance the model's performance across a broader range of subjects.

References

[1] Mohamed Taher Alrefaie. "Arabic Stop Words". In: *Arabic Stop Words* (2022). URL: https://github.com/mohataher/arabic-stop-words/blob/master/README.md.

[2] fatima bendaouch. "Data about Covid 19". In: *Data about Covid 19* (2022). URL: https://github.com/fati-bendaouch/Sentiment-Analysis-for-Moroccan-Dialect/blob/main/README.md.

[3] Mehdi CHEBBAH. "Arabic Stop Words 2". In: *Arabic Stop Words 2* (2021). URL: https://github.com/MehdiCHEBBAH/Analyse-des-sentiments-pour-les-commentaires-arabes/blob/master/ar_stopwords.txt.

[4] Abdeljalil Elouardighi, Mohcine Maghfour, and Hafdalla Hammia. "ElecMorocco2016". In: *International conference on model and data engineering* (2017), pp. 262–274. URL: https://github.com/sentiprojects/ElecMorocco2016/blob/master/README.md.

[5] El Mahdaouy Essefar Ait Baha. "OMCD: Offensive Moroccan Comments Dataset". In: *Language Resources Evaluation (2023)* (2023). URL: https://github.com/kabilessefar/OMCD-Offensive-Moroccan-Comments-Dataset/tree/main.

[6] Kamel Gaanoun et al. "Darijabert: a Step Forward in Nlp for the Written Moroccan Dialect". In: *Darijabert: a Step Forward in Nlp for the Written Moroccan Dialect* (2023).

[7] Moncef Garouani, Hanae Chrita, and Jamal Kharroubi. "stopwords". In: *Digital Technologies and Applications* (2021). Ed. by Saad Motahhir and Badre Bossoufi, pp. 597–608. DOI: 10.1007/978-3-030-73882-2_54.

[8] Moncef Garouani and Jamal Kharroubi. "MAC: An Open and Free Moroccan Arabic Corpus for Sentiment Analysis". In: *Innovations in Smart Cities Applications Volume 5* (2022). URL: https://github.com/LeMGarouani/MAC/blob/main.

[9] Motassim Hamza. "Sentiment Analysis for Darija". In: *Sentiment Analysis for Darija* (2023). URL: https://github.com/hamzaae/DCSA/blob/main/Academic/paper.pdf.

[10] Mourhir Ibrahimi Anass. "Moroccan Darija Offensive Language Detection Dataset". In: *Moroccan Darija Offensive Language Detection Dataset* (2023). URL: https://data.mendeley.com/datasets/2y4m97b7dc/1.

[11] David Liang. "Intro — Getting Started with Text Embeddings: Using BERT". In: *Intro — Getting Started with Text Embeddings: Using BERT* (2024). URL: https://medium.com/@davidlfliang/intro-getting-started-with-text-embeddings-using-bert-9f8c3b98dee6.

[12] "Moroccan Dialect suffixes and prefixes". In: *Moroccan Dialect suffixes and prefixes* (2024). URL: https://tajinequiparle.com/en/moroccan-arabic-grammar.

[13] Errami Mouaad et al. "Sentiment Analysis on Moroccan Dialect based on ML and Social Media Content Detection". In: *International Journal of Advanced Computer Science and Applications* 14 (Apr. 2023), pp. 315–325. DOI: 10.14569/IJACSA.2023.0140347.

[14] Otman Moussaoui and Yacine El Younnoussi. "Pre-training Two BERT-Like Models for Moroccan Dialect: MorRoBERTa and MorrBERT". In: *MENDEL* 29.1 (June 2023), pp. 55–61. DOI: 10.13164/mendel.2023.1.055. URL: https://mendel-journal.org/index.php/mendel/article/view/223.

[15] EL HADRATI Othman. "Darija Sentiment Analysis". In: *Darija Sentiment Analysis* (2024). URL: https://blabla-bdarija.vercel.app.

[16] Sara Ouahabi, El Ouahabi Safâa, and El Wardani Dadi. "Contribution to the Moroccan Darija sentiment analysis in social networks". In: *Social Network Analysis and Mining* 13 (Oct. 2023). DOI: 10.1007/s13278-023-01129-1.

[17] Aissam Outchakoucht and Hamza Es-Samaali. "Moroccan Dialect -Darija- Open Dataset". In: *Moroccan Dialect -Darija- Open Dataset* (2021). arXiv: 2103.09687 [cs.CL]. URL: https://github.com/darija-open-dataset/dataset.

[18] "Sentiment Analysis Dataset in Moroccan Dialect: Bridging the Gap Between Arabic and Latin Scripted dialect". In: *Sentiment Analysis Dataset in Moroccan Dialect: Bridging the Gap Between Arabic and Latin Scripted dialect* (2023). arXiv: 2303.15987 [cs.CL]. URL: https://github.com/MouadJb/MYC/blob/main/README.md.