EduDarijaBERT: An Applied Transformer-Based Framework for Sentiment Analysis in Moroccan Arabic Educational Discourse

Abderrahim Ait Ichou, Ali Ouacha
IPSS Team, Computer Science Department, Faculty of Sciences, Mohammed V University in Rabat, Morocco

***Abstract*—Students' opinions play a pivotal role in the formulation of successful educational policies. However, the computational analysis of the Moroccan Arabic Dialect remains widely viewed as problematic due to a focus on complex morphology and the absence of specific datasets. To close this gap, we propose an end-to-end system for educational sentiment mining, which utilizes EduDarijaBERT—a transformer model explicitly fine-tuned on this academic domain. In the development of the system, we initially constructed and systematically assessed a specifically curated corpus of 12,223 social media comments. To assess the model's reliability, we evaluated the framework on an independent test set. The proposed EduDarijaBERT system achieved an accuracy of 83.15% on this dataset, demonstrating strong generalization capabilities for educational sentiment mining. In addition to these quantitative metrics, the system facilitates a qualitative interpretation of sentiment categories, providing practical insights revealing that administrative challenges are among the major contributors to negative feedback, while active academic support is the key to positive interaction. This study provides universities with a stable, easily implementable system to track student feedback and fuel data-driven innovations.**

Keywords*—Sentiment analysis system; Moroccan Arabic dialect; social media analytics; educational data mining; data engineering framework

I. Introduction

Social media has permanently altered how students share their experiences and feelings about their education online. Sentiment analysis, as a critical application of applied Natural Language Processing (NLP), uncovers these attitudes and behavioral patterns based on spontaneous text posted on platforms such as YouTube, Facebook, and Instagram [1, 2, 3].

Understanding how students navigate the learning process is crucial for institutions striving to improve educational quality and student engagement [1]. In many psychological studies, emotional states—ranging from high anxiety to strong motivation—are viewed as fundamental factors that determine learning outcomes [4]. Therefore, the implementation of sentiment analysis systems serves as an exceptionally practical indicator of institutional performance and examination policies [1,2]. However, transitioning these NLP systems from theoretical models to practical applications remains a significant challenge when dealing with complex Arabic dialects [5, 6].

The Arabic language is known for its rich morphology [7], complicated syntax, and a dominant diglossia between Modern Standard Arabic and its dialects [5, 7]. The predominant dialect in informal digital communication is the Moroccan Dialect (Darija), which lacks a standard orthography and is characterized by frequent code-switching and the use of Latin scripts (Arabizi). This causes a significant bottleneck in data engineering. Moreover, high-quality, domain-specific annotated datasets focused on Moroccan educational discourse are unfortunately scarce, which hinders the implementation of effective institutional monitoring systems [8, 9].

To overcome these bottlenecks, we propose a comprehensive architecture for educational sentiment mining. We designed a dedicated data pipeline to acquire and standardize a novel dataset of 12,223 Moroccan Arabic comments from Facebook, YouTube, and Instagram. To cope with excessive linguistic noise, we employed a mixed semi-automated data processing scheme, which was rigorously controlled by a human-in-the-loop validation process by native speakers to ensure label accuracy.

The core of our proposed framework is EduDarijaBERT, a domain-adapted model checkpoint. An existing base DarijaBERT model was extensively fine-tuned using our educational corpus to develop EduDarijaBERT [10, 11, 12]. This specific variant was trained to capture the nuances of Moroccan students' discourse, including their specific pedagogical and emotional linguistic features. To address inherent data challenges—where positive and negative sentiments are rarely equally distributed—we integrated the SMOTE algorithm to mitigate severe class imbalances. The overall system's best performance was then contrasted with traditional NLP baselines (including TF-IDF/Word2Vec with SVM, KNN, and XGBoost) to justify the need for this domain-adapted transformer architecture.

This study's key contribution lies in its data engineering pipeline and system framework. We provide a targeted pipeline as well as the first high-quality, education-oriented Moroccan Arabic sentiment corpus. On this premise, we explain the development of EduDarijaBERT, which is the main component of our classification architecture. This framework tackles issues such as extreme class imbalance and connects theoretical NLP with educational data mining to provide an automated tool for institutions to analyze student feedback and make necessary policy improvements based on that data.

The rest of this study is organized as follows: Section II discusses related work. The materials and methods are described in Section III, which includes information on data collection, data preprocessing, data annotation, and data cleaning. The results are discussed in detail in Section IV, along with a comparative analysis, the impact of SMOTE, the fine-tuning of EduDarijaBERT, and performance assessment. Section V discusses data analysis and corpus linguistics, which includes an overview of sentiment distribution, statistics, and lexical trends. Lastly, Section VI provides the concluding remarks and future directions.

II. Related Works

This section reviews the existing literature on applied sentiment analysis for Arabic and its dialects on social media. Sentiment analysis has emerged as a critical tool for extracting user opinions from platforms like Twitter, Facebook, and YouTube. To create systems capable of identifying actionable behavioral patterns within this digital landscape, researchers have increasingly deployed computational linguistics, machine learning, and deep learning models.

Several studies have successfully applied sentiment analysis models to the educational sector. For instance, [13] developed an SVM-based sentiment classification system to analyze Arabic course reviews at a Saudi Arabian university, demonstrating how sentiment classification can practically optimize academic services. Similarly, [14] proposed a machine learning model to assess public opinion regarding distance learning in Saudi Arabia. Using the Apache Spark framework alongside regex-based preprocessing pipelines, their Logistic Regression classifier achieved 91% precision on a dataset of Arabic tweets.

Transformer architectures have also become foundational technologies for processing dialectal Arabic. To illustrate, [15] compared several BERT-based models—including AraBERT, QARIB, and AraELECTRA—on a Moroccan dialect sentiment dataset. Their findings emphasized that pre-trained language models can effectively process complex dialectal variations, with QARIB achieving an accuracy of 96%. Addressing real-world data engineering challenges, [16] tackled the issue of data imbalance in Moroccan dialect datasets and evaluated different feature extraction methods (TF-IDF and N-grams). Their results indicated that data resampling significantly enhances system stability, yielding a precision of 90.24%.

Exploring other dialects, [17] compiled a benchmark dataset of 18,589 tweets and proposed a hybrid MARBERT-LSTM model, achieving state-of-the-art accuracy at 91.23%. Furthermore, [18] introduced the AHaSIS shared task to investigate domain-specific sentiment classification within the hospitality industry. By developing an annotated dataset for Saudi and Moroccan dialects, they established a baseline F1-score of 0.81, underscoring the pressing need for domain- and dialect-specific sentiment analysis systems.

To better situate the performance of these architectures, Table I provides a comparison of the most prominent BERT-based models used in Arabic NLP. While models like AraBERT and AraELECTRA rely on massive corpora primarily composed of Modern Standard Arabic (MSA), they often lack the dialectal depth required for Moroccan social media. Models like QARIB and MARBERT improve upon this by including informal tweets and multi-dialectal data. However, DarijaBERT was specifically selected as our base model because it is uniquely tailored to the Moroccan dialect, having been trained on a corpus that specifically captures the nuances of Darija. This specialized focus is critical for our task, as general Arabic models often struggle with the specific morphological and syntactic structures of Moroccan Arabizi and dialect.

TABLE I. Comparison of Pre-Trained Arabic BERT Models

Model	Pre-training Corpus Size	Primary Language/Dialect Coverage
AraBERT (v0.2)	77 GB (~8.8 billion words)	Modern Standard Arabic (MSA)
AraELECTRA	77 GB	Modern Standard Arabic (MSA)
QARIB (mix)	25 GB (122M lines) to 60 GB	MSA and Dialectal Arabic
MARBERT	128 GB (1B Tweets)	Multi-dialectal (General Arabic)
DarijaBERT	691 MB (~3 million sequences)	Moroccan Darija (Arabic script)

Beyond document-level classification, more granular NLP tasks have also been explored. [19] proposed an AraBERT-based Aspect-Based Sentiment Analysis (ABSA) approach that jointly extracts aspect terms and categories. This multi-task method achieved an F1-score of 80.32% and successfully reduced error propagation, facilitating the extraction of highly specific feedback. Additionally, [20] demonstrated that supervised learning can reliably indicate student satisfaction. They concluded that statistical validation (yielding a Cronbach's Alpha of 0.901) is sufficient to ensure the reliability of using sentiment models to evaluate student well-being.

Sentiment analysis has also seen practical deployment in crisis management. [21] implemented a real-time infoveillance system to track Moroccan social media sentiments regarding the COVID-19 pandemic. Their deep learning models provided critical insights into public opinion dynamics, directly supporting public health decision-making. To handle the data engineering complexities inherent in such tasks, [8] created a large-scale Moroccan dialect dataset designed to overcome the challenges posed by Arabizi (the mixture of Arabic and Latin scripts), providing a necessary foundation for representing noisy web text.

Investigating feature engineering, [22] evaluated various machine learning pipelines for Moroccan dialect emotion recognition and found that N-grams paired with TF-IDF weighting significantly improved classifier performance. Building on this, [23] addressed Arabic's complex morphology and syntax using a hybrid deep learning network combining AraBERT, CNN, GRU, and LSTM architectures. Their findings revealed that pairing AraBERT with GRU and Bi-GRU layers achieved an accuracy exceeding 93%. Collectively, these studies highlight the immense potential of advanced NLP models. However, they also underscore the

specific gap our research addresses: there remains a critical need for a fully integrated, domain-adapted system (such as EduDarijaBERT) specifically trained to process highly noisy educational Moroccan Arabizi and deliver actionable insights directly to educational institutions.

While existing literature demonstrates the success of transformer models like QARIB (96% accuracy) or hybrid MARBERT-LSTM architectures in general dialectal tasks, these models often lack the dialectal depth and domain-specific focus required for the noisy pedagogical Arabizi found in Moroccan student discourse. Furthermore, although studies have utilized SMOTE to stabilize imbalanced Moroccan datasets, our analysis suggests that synthetic oversampling can introduce semantic noise that obscures subtle contextual boundaries. Consequently, the EduDarijaBERT architecture was designed to address these specific gaps by employing a human-in-the-loop pipeline for linguistic fidelity and achieving the ability to alleviate class imbalance natively through targeted domain adaptation, rather than relying on artificial data augmentation.

III. Materials and Methods

To develop a successful system for monitoring the online emotions of university students, establishing an effective data acquisition strategy focused on the Moroccan Arabic dialect is essential. We mined interactions from Facebook, Instagram, and YouTube to create our foundational corpus. The targeted data streams primarily focused on computer science, economics, and daily academic activities.

A significant data engineering challenge in this context is Arabizi. To process this Latinized script, we implemented a hybrid pipeline. Initial transliteration into Arabic characters was performed by a Large Language Model (GPT-4), followed by a strict human-in-the-loop review by native speakers to ensure high fidelity and correct spelling errors.

To establish a robust ground truth for our system, we utilized the same language model to generate initial sentiment labels (positive, neutral, and negative), which were subsequently subjected to extensive manual validation by expert annotators. The data sanitization stage entailed the elimination of digital noise, including URLs, platform-specific hashtags, and common stop words. After standardization, light stemming was applied, followed by tokenization to prepare the text for model ingestion.

To train the models and optimize feature definitions, we benchmarked traditional TF-IDF and Word2Vec pipelines against advanced contextual embeddings derived from DarijaBERT. The main architectural product of this process is EduDarijaBERT (Fig. 1). It is a domain-specific adaptation of the original DarijaBERT model, extensively fine-tuned on our custom educational dataset to capture domain-specific linguistic nuances. We compared its performance with that of XGBoost and SVM. SMOTE was employed to balance the dataset, ensuring system robustness for real-world applications characterized by inherent class imbalance. The final system evaluation relied on standard metrics, such as accuracy, precision, recall, and F1-score.

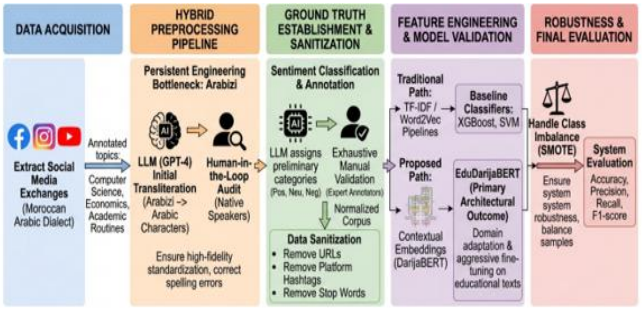

Fig. 1. System architecture of the proposed EduDarijaBERT framework.

A. Data Collection

For this study, we gathered a comprehensive set of user comments across 12,587 pages directly from Facebook, YouTube, and Instagram. One of the key benefits of this corpus is its broad temporal range, spanning 10 years from 2016 to 2025. This longitudinal study allows for the analysis of the linguistic evolution of Moroccan Darija and the students' attitudes through many generations and changing educational contexts. The overall style of these posts was informal, but thematic analysis found that there was a significant emphasis on certain areas of academic study, including mathematics, economics, and computer science. Moreover, many end-users expressed their own experiences about daily life on campus or encouraged the commendation of local universities in international contests. The overall distribution of comments among the three networks was uneven, due to different levels of user engagement across these platforms. This imbalance, however, provided the foundation for our next data engineering pipeline, which was used to create a robust sentiment classification framework that was trained and validated over this comprehensive decade-long dataset, with a high degree of temporal generalizability.

B. Data Preprocessing

1) Text conversion: One of the most difficult parts of the present study was to make the Arabizi data into a form that could be used for quantitative analysis. The corpus consisted of 6,555 comments in Latin-script Moroccan Arabic, which is a dialect with rather frequent French code-switching. This means that it was impossible to perform the computational processing without first establishing orthographic regularity. To overcome this, a two-phase transcription pipeline was implemented. The first phase involved using GPT-4o mini to convert the Latin-script input into Arabic script, which is quite variable. Because automated systems frequently struggle to process complex dialectal features and code-switching, these initial outputs were not considered final. Instead, all conversions were rigorously reviewed by a human team who corrected any vowel and consonant misalignments where the model had failed. The results of this process are illustrated in Table II, which presents a selection of original Latin-script examples alongside their final Arabic equivalents.

TABLE II. EXAMPLES OF COMMENT CONVERSION FROM LATIN SCRIPT TO ARABIC SCRIPT WITH ENGLISH

Latin Script (Franco-Arabic)	Arabic Script (Moroccan Dialect)	English Translation
Bravo madame	برافو أمدام	Bravo, Madam
Merci beaucoup Madame pour cette vidéo	شكراً بزاف أمدام على هاد الفيديو	Thank you very much, Madam, for this video.
Bghina les exercices	بغينا تمارين	We want the exercises.
Merci bien prof	شكراً أستاذة	Thank you very much, Professor.
MERCI Bq chère prof	شكراً بزاف أستاذة	Thank you very much, dear Professor.

C. Data Annotation

The first round of sentiment classification was done using GPT-4o mini, which assigned one of three sentiment classes to every comment. To safeguard analytical rigor and make the manual work on the entire dataset manageable, the total dataset was divided into six different and non-overlapping subsets. The different subsets were assigned to six independent human annotators, each of whom is a native Moroccan dialect speaker. The model's predictions were systematically reviewed by the annotators within their respective partitions, where they corrected incorrect classifications by hand to ensure that the pedagogical and emotional context of the text was properly captured. To address potential algorithmic bias, we analyzed cases where humans overrode the initial labels, as detailed in Table III. For example, GPT-4o mini misidentified polite inquiries (e.g., asking for a WhatsApp number) as negative, which humans corrected to neutral, and religious encouragement (e.g., "Masha'Allah") as neutral, which humans corrected to positive. This correction process ensures the dataset reflects authentic human intent. This human-in-the-loop method was partitioned to ensure an efficient process of creating a large-scale, reliable ground-truth dataset. To assess the reliability of this process, a subset of 200 comments was independently evaluated by three annotators, yielding an overall Fleiss' Kappa of $\kappa = 0.8175$ (95% CI: [0.7375, 0.8975]), which indicates almost perfect inter-annotator agreement. While this agreement subset represents a portion of the total corpus, the narrow confidence interval confirms the statistical robustness and the reliability of the annotation guidelines. An analysis of the disagreement cases revealed that they were primarily driven by "mixed sentiments" and the cultural nuances of Moroccan Darija. In some cases, positive (or negative) feedback on the teacher's speed of presentation was mixed with a positive comment, creating a dilemma for the annotators between a negative, neutral, or positive reading. There were also some differences regarding short snippets of expressions, which were either contextually neutral/factual or interpreted as positive engagement. The examples show that although there is a high level of agreement at the overall level, subjectivity regarding politeness versus dissatisfaction is still problematic in dialectal sentiment analysis. Table IV provides an example of this completed dataset, with the original comments written in Arabic characters and the English translation alongside the approved categories of sentiment.

TABLE III. COMPARISON OF GPT-4O MINI INITIAL LABELING VERSUS HUMAN EXPERT CORRECTION

Human Label	GPT-4o mini Label	Arabic Script (Moroccan Dialect)	English Translation
Neutral	Negative	ما لقيتش رقم الواتساب ديالك	I didn't find your WhatsApp number
Neutral	Negative	مكرهتشي تعدلنا التمارين	I wouldn't mind if you did the exercises for us
Positive	Neutral	تحياتي	My regards / Greetings
Neutral	Positive	غضو البصر يا امة محمد	Lower your gaze, O nation of Muhammad
Positive	Neutral	ما شاء الله وفي ذلك فلتنافس	Masha'Allah, and for that let them compete

TABLE IV. EXAMPLES OF SENTIMENT POLARITY LABELS ALLOTTED TO COMMENTS

Latin Script (Franco-Arabic)	Arabic Script (Moroccan Dialect)	English Translation
Positive	ألف مبروك له ولكل الفائزين، مزيدا من التفوق والنجاح ان شاء الله	A thousand congratulations to him and to all the winners; more excellence and success, God willing.
Negative	النتيجة : هجرة الأدمغة	The result: brain drain.
Neutral	ذهبوا وهم "مغاربة" هل سيرجعون وهم "مغاربة" ؟	They left as "Moroccans"; will they return as "Moroccans"?
Positive	ألف مبروك له ولكل الفائزين، مزيدا من التفوق والنجاح ان شاء الله	A thousand congratulations to him and to all the winners; more excellence and success, God willing.
Negative	النتيجة : هجرة الأدمغة	The result: brain drain.

D. Data Cleaning

The raw corpus underwent extensive cleaning before it became suitable for analysis. The initial step was to convert the text into Arabic script because the comments were originally written in Arabizi, which is Latin-scripted Moroccan Arabic, to ensure data uniformity. We also deleted emojis from 1,055 comments and removed URLs from 775 others. Then there was Arabic normalization. This implied the elimination of typical orthographic inconsistencies: the Alif variants (إأآا) were reduced to a single form, mapping ى to ي, ة to ه, and گ to ك. Even marks of elongation and extended lines such as هههههههه were left out, as they are analytically unproductive. To address the linguistic specificities of the corpus, we implemented a hybrid stop-word removal process by combining the standard NLTK Arabic stop-word list with a manually curated dictionary of Moroccan Darija terms, including dialect-specific pronouns (e.g., أنا, نتا), interrogatives (e.g., واش, فين), and common adverbs (e.g., بزاف, بلا). After tokenization, another 297 entries turned out to be invalid and were dropped. The dataset was reduced to 12,223 useful comments, having started with 12,587 initial comments. The distribution of these collected comments across different social media platforms is shown in Fig. 2.

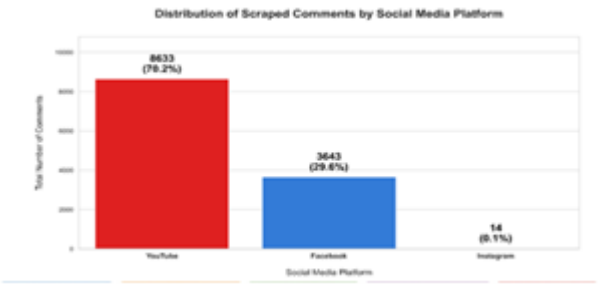

Fig. 2. Distribution of scraped comments by social media platform.

IV. Results and Discussion

This section evaluates the performance of different machine learning algorithms and embedding methods for the sentiment analysis of Moroccan Arabic educational comments. To get a full picture, we evaluated the models using standard classification metrics: accuracy, precision, recall, and F1-score, as well as confusion matrices and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to gain in-depth insights into the models' discriminative power.

A. Comparative Analysis

All experiments showed that transformer-based architectures and ensemble techniques outperformed the traditional linear classifiers. The best performance was achieved when the DarijaBERT [3] model was used with an SVM classifier, with a total accuracy score of 80.17. These results indicate that domain-based pre-trained language models of Moroccan dialect are more effective than generic embeddings at capturing semantic information in Moroccan Dialect.

Traditional machine learning algorithms, including SVM and XGBoost, have a consistent set of results in all samples that used TF-IDF, Word2Vec, and FastText features. TF-IDF provided a good baseline, and its performance was on par with word embeddings. Moreover, a few keywords of the Moroccan dialect therefore exhibit high correlation with sentiment polarity.

B. Impact of SMOTE

To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training partition after performing the train/test split. To prevent data leakage and ensure an unbiased evaluation, the test set remained completely untouched.

Specifically, SMOTE was used to rebalance the training partition of 9,167 samples, generating synthetic instances for the minority classes to yield an expanded training set of 14,484 samples (equally distributed at 4,828 samples per class).

Table VI presents a performance comparison of the classifiers with and without SMOTE. The results indicate that SMOTE improved the recall and F1-score for minority categories (e.g., negative comments), demonstrating enhanced detection of underrepresented sentiments. However, as shown in Table V, this oversampling strategy led to a slight decline in overall accuracy and macro F1-score across several models. This indicates that synthetic oversampling can introduce noise into dialectal Arabic text, where semantic and contextual boundaries are already subtle. Consequently, applying SMOTE presents a clear trade-off: it achieves a more equitable classification for minority opinions at the cost of a slight reduction in global model precision.

TABLE V. Class Distribution Before and After SMOTE Application

Distribution	Negative	Positive	Neutral	Total
Before SMOTE	1284	3055	4828	9167
After SMOTE	4828	4828	4828	14484

TABLE VI. Performance Comparison of Models and Embedding

Embedding / Model	*Method*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
DarijaBERT + SVM	Without SMOTE	0.8017	0.8018	0.8017	0.8009
DarijaBERT + SVM	With SMOTE	0.7856	0.8058	0.7856	0.7901
DarijaBERT + XGBoost	Without SMOTE	0.7987	0.8024	0.7987	0.7973
DarijaBERT + XGBoost	With SMOTE	0.7951	0.8004	0.7951	0.7960
DarijaBERT +Random Forest	Without SMOTE	0.7663	0.7724	0.7663	0.7642
DarijaBERT +Random Forest	With SMOTE	0.7699	0.7839	0.7699	0.7726
DarijaBERT+Logistic Regression	Without SMOTE	0.7837	0.7819	0.7837	0.7827
DarijaBERT +Logistic Regression	With SMOTE	0.7781	0.7833	0.7781	0.7799
DarijaBERT + KNN	Without SMOTE	0.7621	0.7582	0.7621	0.7594
DarijaBERT + KNN	With SMOTE	0.6659	0.7653	0.6659	0.6831
DarijaBERT + Naive Bayes	Without SMOTE	0.5818	0.6818	0.5818	0.6008
DarijaBERT + Naive Bayes	With SMOTE	0.6116	0.6758	0.6116	0.6274
TF-IDF + SVM	Without SMOTE	0.8007	0.8031	0.8007	0.7968
TF-IDF + SVM	With SMOTE	0.7783	0.7867	0.7783	0.7768
TF-IDF + XGBoost	Without SMOTE	0.7579	0.7817	0.7579	0.7584
TF-IDF + XGBoost	With SMOTE	0.7516	0.7799	0.7516	0.7561
TF-IDF + Random Forest	Without SMOT	0.7693	0.7695	0.7693	0.7642

Embedding / Model	*Method*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
	E				
TF-IDF + Random Forest	With SMOTE	0.7575	0.7742	0.7575	0.7684
TF-IDF + Logistic Regression	Without SMOTE	0.7857	0.7845	0.7857	0.7848
TF-IDF + Logistic Regression	With SMOTE	0.7598	0.8017	0.7598	0.7721
TF-IDF + KNN	Without SMOTE	0.6867	0.5846	0.6867	0.5650
TF-IDF + KNN	With SMOTE	0.5949	0.7422	0.5949	0.6137
TF-IDF + Naive Bayes	Without SMOTE	0.7775	0.7806	0.7775	0.7653
TF-IDF + Naive Bayes	With SMOTE	0.7631	0.7950	0.7631	0.7727
FastText + SVM	Without SMOTE	0.7988	0.8011	0.7988	0.7993
FastText + SVM	With SMOTE	0.7853	0.8104	0.7853	0.7912
FastText + XGBoost	Without SMOTE	0.7948	0.7970	0.7948	0.7942
FastText + XGBoost	With SMOTE	0.7883	0.7936	0.7883	0.7899
FastText + Random Forest	Without SMOTE	0.7755	0.7841	0.7755	0.7730
FastText + Random Forest	With SMOTE	0.7673	0.7801	0.7673	0.7695
FastText + Logistic Regression	Without SMOTE	0.7736	0.7730	0.7736	0.7699
FastText + Logistic Regression	With SMOTE	0.7497	0.7797	0.7497	0.7574
FastText + KNN	Without SMOTE	0.7307	0.7382	0.7307	0.7303
FastText + KNN	With SMOTE	0.6787	0.7518	0.6787	0.6959
FastText + Naive Bayes	Without SMOTE	0.4781	0.7119	0.4781	0.5137
FastText + Naive Bayes	With SMOTE	0.5108	0.6894	0.5108	0.5438
Word2Vec + SVM	Without SMOTE	0.7893	0.7869	0.7893	0.7873
Word2Vec + SVM	With SMOTE	0.7723	0.7916	0.7723	0.7778
Word2Vec + XGBoost	Without SMOTE	0.7929	0.7912	0.7929	0.7916
Word2Vec + XGBoost	With SMOTE	0.7919	0.7958	0.7919	0.7935
Word2Vec + Random Forest	Without SMOTE	0.7935	0.7945	0.7935	0.7924
Word2Vec + Random Forest	With SMOTE	0.7860	0.7927	0.7860	0.7884
Word2Vec + Logistic Regression	Without SMOTE	0.7778	0.7750	0.7778	0.7755
Word2Vec + Logistic Regression	With SMOTE	0.7575	0.7764	0.7575	0.7629
Word2Vec + KNN	Without SMOTE	0.7664	0.7667	0.7664	0.7659
Word2Vec + KNN	With SMOTE	0.7268	0.7798	0.7268	0.7379
Word2Vec + Naive Bayes	Without SMOTE	0.6931	0.7489	0.6931	0.7039
Word2Vec + Naive Bayes	With SMOTE	0.6787	0.7337	0.6787	0.6898

C. Discussion of findings

The robust performance of SVM across almost all embedding types (achieving approximately 80% accuracy) suggests that the sentiment boundaries in Moroccan Arabic social media comments are relatively well-defined when projected into high-dimensional spaces.

However, Naive Bayes and KNN showed significant sensitivity to the feature extraction method. For instance, Naive Bayes performed poorly with Word2Vec and FastText (often falling below 70%) but showed a substantial improvement with TF-IDF (up to 77.7%). This indicates that for dialectal Arabic, the presence of specific tokens (captured by TF-IDF) is more informative for probabilistic models than the semantic clusters formed by dense embeddings.

1) Impact of fine-tuning EduDarijaBERT on sentiment classification:

a) Experimental setup and hyperparameter configuration: For the fine-tuning experiments, sentence representations were extracted from the last hidden layer of the fine-tuned EduDarijaBERT model using the [CLS] token. A maximum sequence length of 128 tokens was utilized to tokenize the input streams. All inputs were padded and truncated to ensure a consistent input size across the dataset. The model was fine-tuned for five epochs with a training and evaluation batch size of 16. Optimization was performed using the AdamW optimizer with the default learning rate and a weight decay of 0.02. To address the inherent class imbalance, a weighted cross-entropy loss function was employed using dynamically calculated class weights. Model checkpoints were saved and evaluated at the end of every epoch, and the Area Under the Receiver Operating Characteristic Curve (AUC-

ROC) was utilized as the primary metric to select the best-performing model.

Subsequently, a Support Vector Machine (SVM) classifier employing a Radial Basis Function (RBF) kernel was trained using the extracted embeddings. The regularization parameter was set to 1.0, and the kernel coefficient was set to 'scale'. To facilitate class-wise performance analysis, probability estimation was enabled. The random state was fixed at 42 to guarantee reproducibility. All experiments were executed on a GPU (CUDA) when available, falling back to a CPU otherwise. Table VII summarizes the complete hyperparameter configuration.

TABLE VII. HYPERPARAMETERS OF THE FINE-TUNED EDUDARIJABERT + SVM MODEL

Component	*Hyperparameter*	*Value*
EduDarijaBERT	Pretrained model	DarijaBERT (SI2M-Lab/DarijaBERT)
	Max sequence length	128
	Padding	True
	Truncation	True
	Ignore mismatched sizes	True
Fine-tuning Configuration	Number of epochs	5
	Train batch size	16
	Eval batch size	16
	learning rate	Default (5e-5 from AdamW)
	Weight decay	0.02
	Warmup steps	300
	Optimizer	AdamW (default)
	Loss function	CrossEntropyLoss (weighted)
	Class weighting	Balanced (computed)
	Evaluation strategy	Per epoch
	Save strategy	Per epoch
	Metric for best model	AUC
	Device	GPU (CUDA) / CPU
SVM classifier	Kernel	RBF
	Regularization parameter (C)	1.0
	Gamma	scale
	Probability estimates	Enabled
	Random state	42

In a subsequent experiment, EduDarijaBERT was fine-tuned on our Moroccan Arabic educational dataset before employing the SVM classifier. While DarijaBERT was used as a frozen feature extractor in previous experiments, fine-tuning enables the model to learn internal representations that are sensitive to the linguistic, pedagogical, and emotional nuances of the Moroccan Arabic educational discourse.

The fine-tuned EduDarijaBERT + SVM model achieved the best score of 83.15% and was the best performer across all settings. This model significantly outperformed the non-fine-tuned DarijaBERT + SVM setting (which achieved an accuracy of 80.17%), thus underscoring the critical importance of domain adaptation for low-resource dialects like Moroccan Arabic.

Table VIII illustrates that fine-tuning improved performance across all evaluation metrics, especially recall and F1-score. This indicates better generalization of the fine-tuned model across sentiment categories. This is particularly evident for the negative class, which is often challenging to classify due to class imbalance and linguistic diversity.

TABLE VIII. PERFORMANCE OF FINE-TUNED EDUDARIJABERT + SVM

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
EduDarijaBERT + SVM (Fine-tuned)	0.8315	0.8299	0.8315	0.8305

b) Overall discussion and implications: The experiments demonstrate that using classical machine learning classifiers and sparse representations such as TF-IDF provides a strong baseline for sentiment classification in Moroccan Arabic text, especially when paired with robust classifiers such as SVM. These findings confirm that lexical cues remain highly discriminative in dialectal Arabic.

Nonetheless, significant improvements and increased robustness across all sentiment classes were achieved when we fine-tuned EduDarijaBERT – a domain-adapted transformer model – through additional training on Moroccan Arabic educational data. This improvement can be attributed to the model's ability to capture linguistic nuances, colloquialisms, and sentiment cues within the educational domain.

Overall, although traditional approaches offer competitive speed and accuracy, the fine-tuned transformer models demonstrated superior performance for educational monitoring and feedback analysis.

2) Comparative performance analysis: The performance differences observed in the experiment with SMOTE-balanced data were minimal, as illustrated in Fig. 3. Across all evaluation metrics, the SVM classifier combined with DarijaBERT embeddings achieved robust results. Without SMOTE, the model reached an accuracy of 80.2%, whereas with SMOTE, the accuracy decreased slightly to 78.6%. When comparing precision, recall, and F1-score, only marginal gains were observed with SMOTE, with precision increasing from 0.802 to 0.806. However, after applying SMOTE, recall declined to 0.786, and the F-score remained comparable in both settings. These findings suggest that either SMOTE enhances the model's ability to discriminate the minority class, or conversely, that DarijaBERT embeddings already provide strong discriminative capability on the imbalanced dataset.

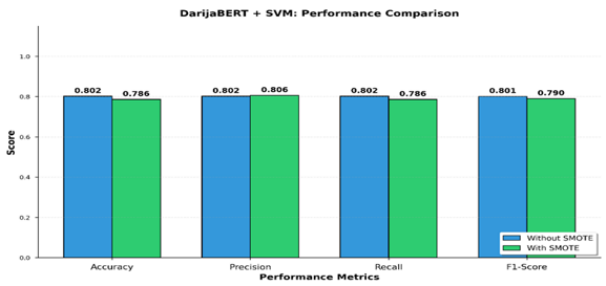

Fig. 3. Comparative performance metrics of DarijaBERT+SVM with and without SMOTE.

a) Confusion matrix analysis: Fig. 4 presents the confusion matrices of DarijaBERT+SVM in all the experimental settings, which provide detailed insights into the classification patterns. Without SMOTE, the model performs well on the majority positive class (1,405 correct predictions out of 1,612), but it is evident that there is confusion between the other categories. Specifically, 179 neutral samples are misclassified as positive, and 131 negative samples are misclassified as neutral; this is indicative of class imbalance.

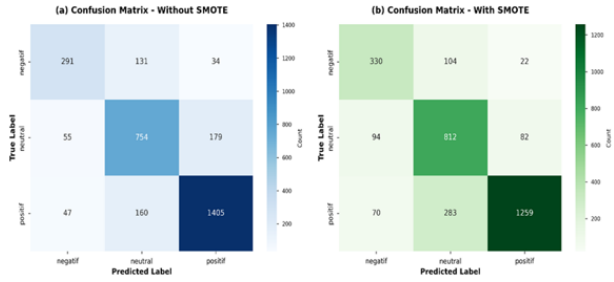

Fig. 4. Confusion matrices for DarijaBERT+SVM classification (a) Without SMOTE and (b) With SMOTE.

By applying SMOTE, the proper classification of minority classes is drastically enhanced. True positives in the negative class improve from 291 to 330, representing an improvement of 13.4%. In the case of the neutral class, the number of correct predictions increases from 754 to 812, which is an increment of 7.7%. However, this improvement comes at the cost of a larger number of positive instances falsely classed as neutral, increasing from 160 to 283. This result illustrates a classic trade-off in cost-sensitive classification between the recall of minority classes and the accuracy of the majority class.

b) Receiver operating characteristic analysis: The Receiver Operating Characteristic (ROC) curves are presented in Fig. 5. The combination of DarijaBERT with an SVM classifier demonstrates strong discriminative capability across each sentiment level individually. In the configuration without SMOTE (Fig. 5a), the weighted average AUC achieves a robust score of 0.917. The scores of the categories varied: the neutral class recorded the lowest AUC at 0.889, while the positive class achieved the highest AUC at 0.931 (with the negative class at 0.929). It is not surprising that the positive class exhibited the best overall ROC profile, as its prevalence in the raw training distribution allows the model to characterize it more effectively.

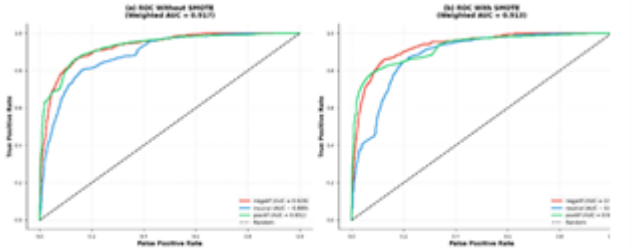

Fig. 5. ROC curves and AUC scores for multi-class sentiment classification (a) Without SMOTE and (b) With SMOTE.

When SMOTE is utilized (Fig. 5b), the weighted average AUC remains highly competitive at 0.913, and the AUC of the negative class increases to 0.932. This demonstrates that the model's ability to identify the minority class is improved. The minor decreases in AUC for the neutral class (from 0.889 to 0.885) and the positive class (from 0.931 to 0.924) reflect the distribution shift caused by synthetic oversampling; specifically, expanding the decision space of the minority categories slightly reduces the classification margins of the majority classes. Regardless of the balancing approaches, the AUC results show that DarijaBERT embeddings capture rich semantic content and can, therefore, be used to classify Moroccan Arabic sentiments in both settings (AUC > 0.88 across all classes).

c) Detailed performance metrics: Table IX shows the effect of SMOTE on all classes, with the negative class experiencing the most significant impact. SMOTE improved the recall from 63.8% to 72.4%, alongside a reduction in precision from 74.0% to 66.8%. This demonstrates that SMOTE effectively reduces false negatives for the minority class at the expense of an increase in false positives, as theoretically anticipated by oversampling processes. The neutral class exhibits a similar trend, with an increase in recall (+7.7%) and a decrease in precision (-6.2%). For the majority positive class, the SMOTE-adapted decision boundary results in a more conservative decision rule, leading to a gain in precision (+6.4%) but a noticeable decrease in recall (-10.4%).

At the macro-average level, SMOTE yields an increase in recall (+2.4%) accompanied by a decrease in precision (-2.7%), resulting in a slight overall decrease in the F1-score (-0.7%). Weighted average precision shows a marginal improvement (+0.5%), as the precision gains in the heavily represented majority class outweigh the precision losses in the minority classes. These findings suggest that the choice of a balancing strategy must be guided by the nature of the application: if the primary goal is maximizing minority class identification (e.g., detecting at-risk students experiencing negative sentiment), SMOTE is highly beneficial. Conversely, for precision-critical applications, maintaining the original, natural class distribution may be preferable.

TABLE IX. COMPREHENSIVE PERFORMANCE COMPARISON OF DARIJABERT+SVM WITH AND WITHOUT SMOTE

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
Negative	Precision	0.740	0.668	-9.7
	Recall	0.638	0.724	+13.5
	F1-Score	0.686	0.695	+1.3
	Support	456	456	-
Neutral	Precision	0.722	0.677	-6.2
	Recall	0.763	0.822	+7.7
	F1-Score	0.742	0.743	+0.1
	Support	988	988	-
Positive	Precision	0.868	0.924	+6.4
	Recall	0.872	0.781	-10.4
	F1-Score	0.870	0.846	-2.8
	Support	1612	1612	-
Overall	Accuracy	0.802	0.786	-2.0
	Macro Avg Precision	0.777	0.756	-2.7
	Macro Avg Recall	0.758	0.776	+2.4
	Macro Avg F1-Score	0.766	0.761	-0.7
	Weighted Avg Precision	0.802	0.806	+0.5
	Weighted Avg Recall	0.802	0.786	-2.0
	Weighted Avg F1-Score	0.801	0.790	-1.4

d) Fine-tuned EduDarijaBERT performance: The confusion matrix for EduDarijaBERT is presented in Fig. 6. Fine-tuning the model results in 320 correctly predicted negative instances (achieving a recall of 70.2%), 763 correctly classified neutral instances (77.2% recall), and 1,458 correctly identified positive instances (90.4% recall). Overall, performance improves across all sentiment classes. The misclassifications between extreme polarities also decrease: there are exactly 39 positive-to-negative and 39 negative-to-positive misclassifications, showing a clear improvement over the baseline setting. Furthermore, confusion between the neutral class and the other sentiment categories is also reduced. Specifically, the neutral category achieves 763 correct predictions, with only 77 and 148 cases being misclassified as negative and positive, respectively. These findings indicate that the fine-tuning process successfully captures domain-specific sentiment cues and the contextual attributes of the educational Moroccan Arabic discourse.

The ROC curves in Fig. 7 demonstrate the strong discriminative performance of the EduDarijaBERT model, achieving a weighted AUC of 0.909. This score is slightly lower than the base DarijaBERT model, but it still has a very strong profile. The positive sentiment class has the best AUC (0.927), followed by the negative sentiment class (0.916) and the neutral sentiment class (0.877). Also, the AUC interval for the best and worst classes for EduDarijaBERT is 0.050, while the interval for the baseline was 0.042 for DarijaBERT without SMOTE. Therefore, this model suits operational settings better where there is a more consistent distribution of classification errors for each classification category.

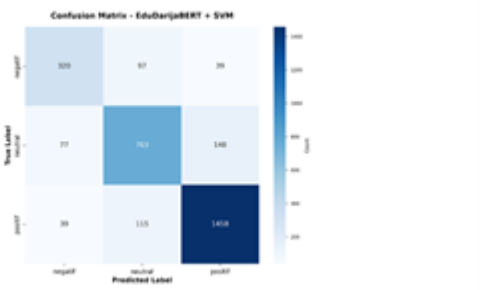

Fig. 6. Confusion matrix of fine-tuned EduDarijaBERT + SVM.

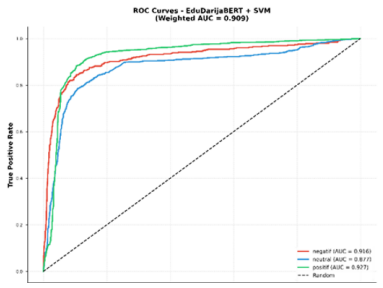

Fig. 7. ROC curves for fine-tuned EduDarijaBERT + SVM.

TABLE X. COMPREHENSIVE PERFORMANCE COMPARISON BETWEEN DARIJABERT AND EDUDARIJABERT IN EDUCATIONAL SENTIMENT ANALYSIS

Metric	*DarijaBERT (Best)*	*EduDarijaBERT*	*Improvement*
Accuracy	0.802	0.832	+3.0%
Precision	0.802	0.830	+2.8%
Recall	0.802	0.832	+3.0%
F1-Score	0.801	0.831	+3.0%
Macro Avg Precision	0.777	0.801	+2.4%
Macro Avg Recall	0.758	0.793	+3.5%
Macro Avg F1-Score	0.766	0.797	+3.1%
Negative Precision	0.740	0.732	-0.8%
Negative Recall	0.724 (SMOTE)	0.706	-1.8%
Negative F1-Score	0.695 (SMOTE)	0.719	+2.4%
Neutral Precision	0.722	0.784	+6.2%
Neutral Recall	0.822 (SMOTE)	0.765	-5.7%
Neutral F1-Score	0.743 (SMOTE)	0.775	+3.2%

Metric	DarijaBERT (Best)	EduDarijaBERT	Improvement
Positive Precision	0.924 (SMOTE)	0.886	-3.8%
Positive Recall	0.872	0.908	+3.6%
Positive F1-Score	0.870	0.896	+2.6%

Table X shows a detailed overview of the performance comparison between the baseline DarijaBERT and the fine-tuned model, EduDarijaBERT. The overall results show a significant improvement across the main metrics, including a 3.1% improvement in the macro-averaged F1-score. Additionally, there were significant gains for specific classes, with the neutral class improving by 6.2% in precision and the positive class by 3.6% in recall. Overall improvements are, however, accompanied by some minor sacrifices in some of the class-specific metrics, which show a small decrease.

e) Robustness check and cross-validation analysis: To evaluate its real-world generalization capability, the proposed model was evaluated on an independent 3,056-sample test set. Tables XI and XII present the detailed per-class performance breakdown. To further evaluate its real-world generalization capability beyond cross-validation, the proposed model was evaluated on an independent 3,056-sample test set. Tables XII and XIII present the detailed per-class performance breakdown.

TABLE XI. EDUDARIJABERT CLASSIFICATION REPORT ON THE INDEPENDENT TEST SET

Model	Precision	Recall	F1-Score	Support
Negative	0.7339	0.7018	0.7175	456
Neutral	0.7826	0.7723	0.7774	988
Positive	0.8863	0.9045	0.8953	1612
Accuracy			0.8315	3056
Macro Avg	0.8009	0.7928	0.7967	3056

TABLE XII. DARIJABERT CLASSIFICATION REPORT ON THE INDEPENDENT TEST SET

Model	Accuracy (Mean ± SD)	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)
Negative	0.7405	0.6382	0.6855	456
Neutral	0.7215	0.7632	0.7418	988
Positive	0.8684	0.8716	0.8700	1612
Accuracy			0.8017	3056
Macro Avg	0.7768	0.7576	0.7657	3056

The overall accuracy was 83.15% for EduDarijaBERT, compared to 80.17% for the baseline model, on the independent test set. Interestingly, there was no balancing using SMOTE in the proposed framework, and yet it retained its performance, while the baseline framework's performance dropped to 78.57% when using SMOTE balancing. The domain-adapted model performance profile was well balanced, with the macro-average F1-score improving from 0.766 to 0.797. Specifically, the F1-score increased significantly to 0.777 for the neutral class and 0.718 for the minority negative class, showing a clear improvement over the baseline. Additionally, EduDarijaBERT demonstrated strong discriminative performance within each class (AUC scores were tightly grouped from 0.877 to 0.927) and remained stable. The model is able to alleviate class imbalance natively with high-quality performance on underrepresented classes without the need for synthetic oversampling. These comparative results demonstrate that targeted fine-tuning on educational Moroccan Arabic data achieves better and more stable classification results for the automatic monitoring of students' sentiments.

The independent test set accuracy of 83.15% provides a conservative and realistic characterization of the system's performance on unseen data. This result reflects the model's real-world generalization capability. The performance level is influenced by the temporal drift and linguistic evolution inherent in a dataset spanning a full decade, from 2016 to 2025. Furthermore, this evaluation confirms the absence of data leakage, as the independent test set remained completely untouched during the SMOTE application and training phases to ensure an unbiased assessment.

V. DATA ANALYSIS AND CORPUS LINGUISTICS

The statistical characteristics of student discourse and its sentiment distribution were examined across the 12,223 manually verified comments. This is a large-scale study that enables a comprehensive grasp of learners' opinions in the context of this specific feedback in Morocco.

A. Sentiment Distribution

The overall tone of the corpus is positive, reflecting a high level of student engagement and satisfaction. However, the negative subset is still needed because it provides detailed information on the educational problems faced by learners. This balanced view is essential to generate a model that reflects the characteristics of the Moroccan language medium (Darija) in an academic context (Fig. 8).

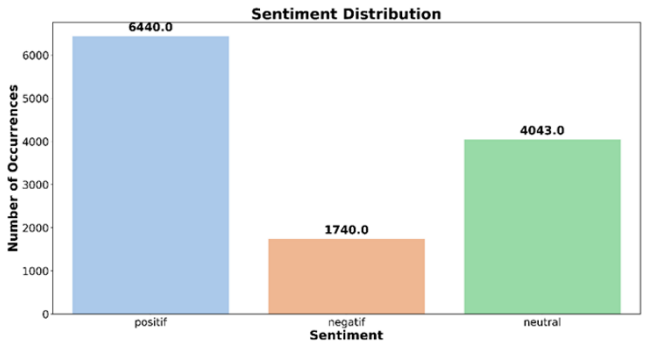

Fig. 8. Sentiment distribution within the Moroccan Arabic educational corpus.

B. Statistical Summary

The students' expressions of the sentiment classes are not equally simple: positive comments are shorter and more frequent, whereas negative ones are more complex, with a maximum length of 522 words. This suggests that when

students write negative comments, they provide a more detailed justification than when they give positive comments.

TABLE XIII. QUANTITATIVE ANALYSIS AND LEXICAL DIVERSITY OF THE CORPUS

Sentiment Label	*Total Number of Words*	*Number of Distinct Words*	*Max Sentence Length*
Positive	39258	9158	233
Negative	22224	8240	522
Neutral	32651	9571	306

C. Lexical Trend Analysis

The most common unique words offer a detailed view of the topics and vocabulary used in the three sentiment classes. Following the overview is a qualitative description which is supplemented by a word frequency distribution, outlining the main concerns and communication styles of the students.

Positive Sentiment: As depicted in Fig. 9, the discourse is characterized by a strong sense of gratitude and religious invocation. Words like "شكرا" (Thank you) and "الله" (God) are used most frequently, reflecting a cultural tendency to express appreciation for educational material or mentors through traditional blessings.

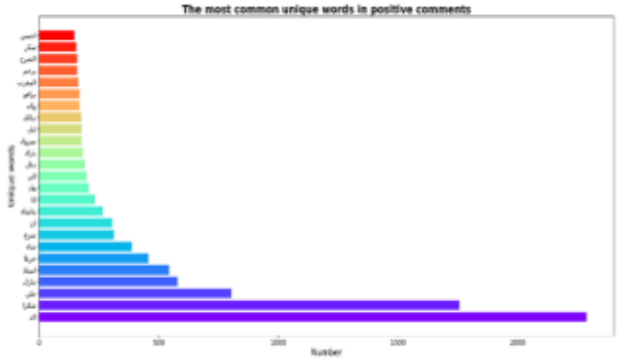

Fig. 9. Top 25 most frequent unique words in positive comments

Neutral Sentiment: The neutral comments are mostly cooperative and inquisitive, as seen in Fig. 10. The use of the words (Teacher), (I/Me), and (Of/Belonging to) suggests that students use a neutral tone mainly to ask for clarifications and to introduce themselves or to discuss administrative and educational course matters.

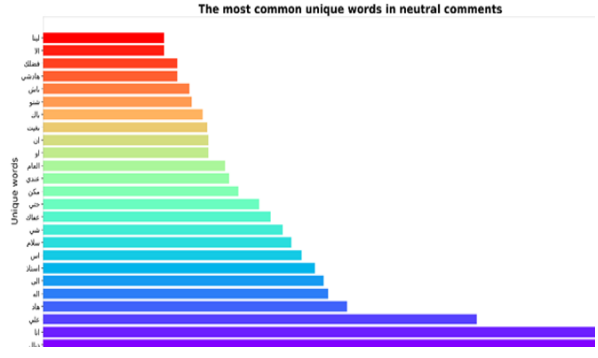

Fig. 10. Top 25 most frequent unique words in neutral comments.

Negative Sentiment: As seen in Fig. 11, the lexicon becomes more critical, shifting from hopeful to sociopolitical. France, Morocco, and Education provide pointers to deep discussions about language policies, educational reforms, student mobility, and migration. Furthermore, administrative and institutional challenges emerge as significant contributors to negative feedback, as students frequently express frustration regarding academic procedures and bureaucratic hurdles within the educational system.

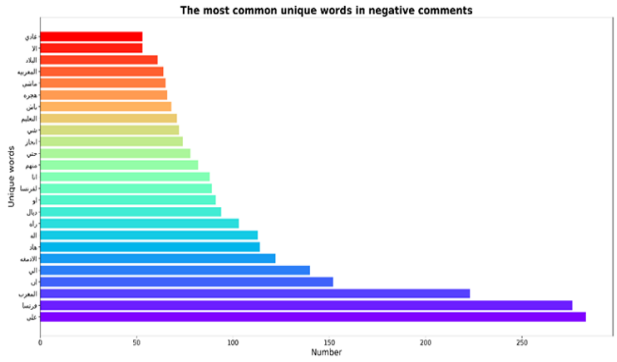

Fig. 11. Top 25 most frequent unique words in negative comments.

VI. CONCLUSION

This study aimed to propose and evaluate a practical model for monitoring the online sentiments of Moroccan university students. The linguistic analysis demonstrated distinct behavioral patterns associated with positive and negative sentiments: expressions of gratitude and religious invocations were correlated with positive feelings, whereas rigid institutional rules and tensions with the broader sociopolitical environment were associated with negative ones. Mapping this dichotomy provides a nuanced understanding of how this demographic navigates its academic reality.

The domain-specific fine-tuning of DarijaBERT with educational corpora to develop the proposed EduDarijaBERT demonstrated a substantial operational advantage. The customized system outperformed the general-purpose baseline in terms of both accuracy and the precision-recall curve. Crucially, EduDarijaBERT effectively classified the minority class of negative sentiment—traditionally a challenging task in Natural Language Processing (NLP)—without relying on synthetic data augmentation techniques such as SMOTE. The underlying principle of this study underscores a core tenet of applied NLP: domain adaptation is vital for successfully classifying local, dialectal text in real-world applications.

Future work will focus on expanding the training dataset to include other Maghrebi dialects, thereby broadening the geographical applicability of the framework. Furthermore, transitioning from static data analysis to longitudinal emotion monitoring offers a promising avenue for real-time institutional observation. Additionally, exploring a formal teacher-student knowledge distillation approach represents a compelling pathway for deployment in resource-constrained production environments. Ultimately, this framework is not only ready for implementation within educational institutions but can also be

scaled to gather actionable insights from large-scale student feedback across Morocco.

REFERENCES

[1] Asok, C., & M, R. (2018). Analyzing and mining social media data for predicting student's learning experiences. IJERT. https://doi.org/10.17577/IJERTCONV4IS17041.

[2] Berea, A., Rand, W., Wittmer, K., University of Maryland, Gerard Wall, & vibeffect. (2015). Social media analysis for higher education. In Washington Academy of Sciences. https://www.washacadsci.org/Journal/Journalarticles/V.101-3-Social%20Media%20Analysis%20for%20Higher%20Education%20A.%20Berea%20et%20al.pdf.

[3] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012. doi: 10.2200/S00416ED1V01Y201204HLT016.

[4] Pekrun, Reinhard. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. Educational Psychology Review. 18. 315-341. 10.1007/s10648-006-9029-9.

[5] Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2018b). A comprehensive survey of arabic sentiment analysis. Information Processing & Management, 56(2), 320–342. https://doi.org/10.1016/j.ipm.2018.07.006.

[6] Shaalan, Khaled. (2010). Nizar Y. Habash, Introduction to Arabic natural language pro-cessing (Synthesis lectures on human language technologies). Machine Translation. 24. 285-289. 10.1007/s10590-011-9087-8.

[7] Farghaly, Ali & Farghaly, Ali & Shaalan, Khaled & Khaled,. (2009). Arabic Natural Language Processing: Challenges and Solu-tions. ACM Transactions on Asian Language Information Processing (TALIP). 8. 1-.

[8] M. Jbel, M. Jabrane, I. Hafidi, and A. Metrane, "Sentiment Analysis Dataset in Moroccan Dialect: Bridging the Gap Between Arabic and Latin Scripted dialect," Research Paper, 2024.

[9] Mihi, Soukaina & Ait Benali, Brahim & EL, Ismail & Arezki, Sara & Laachfoubi, Nabil. (2020). MSTD: Moroccan Sentiment Twitter Dataset. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0111045.

[10] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

[11] Gaanoun, Kamel & Naira, Abdou Mohamed & Allak, Anass & Benelallam, Imade. (2023). Darijabert: a Step Forward in Nlp for the Written Moroccan Dialect. 10.21203/rs.3.rs-2560653/v1.

[12] Antoun, Wissam & Baly, Fady & Hajj, Hazem. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding.

[13] Louati, H. Louati, E. Kariri, F. Alaskar, and A. Alotaibi, "Sentiment Analysis of Arabic Course Reviews of a Saudi University Using Support Vector Machine," Applied Sciences, vol. 13, no. 23, p. 12539, 2023.

[14] Almalki, Jameel. (2022). A machine learning-based approach for sentiment analysis on distance learning from Arabic Tweets. PeerJ Computer Science. 8. e1047. 10.7717/peerj-cs.1047.

[15] Bourahouat, Ghizlane & Abourezq, Manar & Najima, Daoudi. (2023). Improvement of Moroccan Dialect Sentiment Analysis Using Arabic BERT-Based Models. Journal of Computer Science. 20. 157-167. 10.3844/jcssp.2024.157.167.

[16] Nassr, Z., Benabbou, F., Sael, N., & Hamim, T. (2025). Improving sentiment analysis performance on imbalanced Moroccan dialect datasets using resample and feature extraction techniques. Information, 16(1), 39. https://doi.org/10.3390/info16010039.

[17] L. Ouchene and S. Bessou, "Benchmarking sentiment analysis of algerian arabic dialect on X (Twitter)," Journal of King Saud University - Computer and Information Sciences, vol. 37, p. 358, 2025.

[18] M. Alharbi et al., "AHASIS: Shared Task on Sentiment Analysis for Arabic Dialects," arXiv preprint arXiv:2511.13335, 2025.

[19] Fadel, M. Saleh, R. Salama, and O. Abulnaja, "MTL-AraBERT: An Enhanced Multi-Task Learning Model for Arabic As-pect-Based Sentiment Analysis," Computers, vol. 13, no. 4, p. 98, 2024.

[20] O. Chamorro-Atalaya et al., "Quadratic vector support machine algorithm, applied to prediction of university student satisfac-tion," Indonesian Journal of Electrical Engineering and Computer Science, vol. 27, no. 1, pp. 139-148, 2022.

[21] Ghanem et al., "Real-Time Infoveillance of Moroccan Social Media Users' Sentiments towards the COVID-19 Pandemic and Its Management," International Journal of Environmental Research and Public Health, vol. 18, no. 22, p. 12172, 2021.

[22] M. Errami et al., "Sentiment Analysis on Moroccan Dialect based on ML and Social Media Content Detection," International Journal of Advanced Computer Science and Applications, vol. 14, no. 3, 2023.

[23] Ouza, A., Ouacha, A., Rachidi, A., Ghmary, M. E., & Choukri, A. (2024). Enhancing Arabic sentiment analysis using ARABERT and deep learning models. Studies in Computational Intelligence. https://doi.org/10.1007/978-3-031-65038-3_15.