Advancing Sentiment Analysis for Colloquial Arabic: A Comparison of Machine Learning and Lexicon-Based Approaches for the Palestinian Dialect

Khalid Rabaya [1], Ahmad Hasasneh [2,*], and Khalil Rantisi [3]

[1] Department of Computer Science, Faculty of Information Technology, Arab American University, Jenin, Palestine
[2] Department of Natural, Engineering, and Technology Sciences, Faculty of Graduate Studies, Arab American University, Ramallah, Palestine
[3] Department of Computer Science, Faculty of Information Technology, Al-Quds Open University, Ramallah, Palestine
Email: khalid.rabayah@aaup.edu (K.R.); ahmad.hasasneh@aaup.edu (A.H.); khalil.rantisi@aaup.edu (K.Ran.)
*Corresponding author

***Abstract*—Formal Arabic is less commonly used in everyday communication, especially on social media, where colloquial Arabic dominates. However, most Arabic Natural Language Processing (NLP) tools, including sentiment analysis systems, are primarily developed for Modern Standard Arabic (MSA), making it difficult to analyze dialectal content. This study addresses this gap by focusing on the Palestinian dialect, which is widely used across the Levant region. Two main approaches were explored: machine learning and lexicon-based methods. A manually labeled dataset of Facebook telecom service reviews was collected, cleaned, and split into training and testing sets. Five machine learning algorithms and a neural network model (Multilayer Perceptron) were evaluated, while the lexicon-based approach used a predefined sentiment lexicon for colloquial Arabic. The machine learning models outperformed the lexicon-based method, achieving accuracy scores between 92.2% and 94.0%, with Naïve Bayes yielding the highest F1-Score (95.0%). The neural network model achieved 95.0% in both accuracy and F1-Score, while the lexicon-based method scored 81.0% accuracy and a 67.0% F1-Score. Although no hybrid model was implemented, the findings suggest future research could explore combining lexical and machine learning approaches.**

***Keywords*—language, social media, colloquial Arabic, machine learning, lexicon-based approach, sentiment analysis, multilayer perceptron classifier**

I. Introduction

Sentiment analysis is one of the major branches of Natural Language Processing (NLP), which involves the detection of emotions, being positive, negative, or neutral, embedded in a text. This technique is gaining considerable traction in fields such as marketing and customer analysis, driven by the widespread adoption of social media platforms such as Facebook, which has over 3.14 billion monthly active users [1]. These platforms produce massive user-generated data, enabling sentiment analysis to effectively gauge public opinion on specific topics or services. Applying sentiment analysis to Arabic texts faces significant challenges due to the diversity of colloquial dialects, which differ distinctly from Modern Standard Arabic (MSA) [2]. Natural Language Processing (NLP) tools and sentiment models designed for MSA often fail to process colloquial dialects effectively due to their linguistic variations [2]. This research gap motivates us to develop a tailored sentiment analysis model for the Palestinian dialect, evaluating its effectiveness in classifying sentiments expressed in colloquial Arabic.

Formal Arabic is only used in official contexts, such as government and education, but daily communication among people is done using slang Arabic, especially on social media platforms. The Palestinian dialect, prevalent in Palestine, Jordan, Lebanon, and Syria, poses significant challenges for sentiment analysis due to its unique linguistic features. This dialect includes unique phonological, lexical, and syntactic characteristics that are not present in MSA. As a result, lexicon-based sentiment tools, which rely on predefined lists of phrases and associated sentiment, often struggle to accurately interpret these dialectal variations [3, 4]. For example, the Arabic word (زفت) literally means "asphalt" while it carries a strong negative impression about an event or item, cannot be found in formal lexicons. Deep learning-based sentiment analysis, leveraging neural networks, offers a powerful approach for processing colloquial Arabic [5]. This study evaluates the accuracy of deep learning algorithms in classifying sentiments in colloquial Arabic [5].

This study aims to develop and compare sentiment analysis models using machine learning approaches, including a Multilayer Perception (MLP), and lexicon-based methods. Using a dataset of customer

Manuscript received May 19, 2025; revised June 30, 2025; accepted August 12, 2025; published November 7, 2025.

reviews from PalTel's Facebook page, preprocessed and manually annotated by native speakers and reviewed by linguistic experts, we aim to enhance sentiment classification accuracy and contribute insights for NLP research on other Arabic dialects. Building on studies such as Refs. [6, 7], this work contributes to Arabic sentiment analysis by addressing the complexities of dialectal expressions.

While sentiment analysis has been widely applied to Arabic dialects such as Egyptian, Levantine, and Gulf, the Palestinian dialect remains significantly underrepresented in existing research, despite its active use on social media platforms across the Levant region. This study introduces several key contributions: it presents a large, expert-annotated dataset of 8895 Facebook reviews written in the Palestinian dialect; develops a sentiment lexicon with 7027 dialect-specific slang terms, conducts a comparative analysis of machine learning, neural network, and lexicon-based approaches and explores the impact of class imbalance on model performance.

The paper is organized as follows: Section II reviews the literature on Arabic sentiment analysis, highlighting recent advances and key challenges. Section III details the methodology, including data collection, preprocessing, and model development. Section IV evaluates existing sentiment analysis models. Section V discusses the obtained results, and Section VI concludes with findings and future research directions.

II. Literature Review

A growing body of work in Arabic sentiment analysis has significantly improved sentiment classification, particularly in dialectal contexts such as social media and user reviews, which motivates our interest in the unique expressions found in the Palestinian dialect. Recent studies [8–15] offer meaningful insights into the mechanisms of Arabic sentiment analysis and attempt to tackle several pressing challenges in this evolving field.

Research in Arabic sentiment analysis has explored diverse techniques, from lexicon-based dictionaries to neural networks, helping shape our own approach to address the Palestinian dialect's complexities. Some studies have adopted CLASENTI-based frameworks [9], while others have proposed machine and deep learning models [10, 16–18]. Lexicon-based methods [8, 12–15] and emoji-based models [11, 19] have been tested on datasets like tweets and reviews, motivating us to adapt some of these for our Paltel Facebook reviews.

Recently, there has been growing interest in utilizing emojis to improve sentiment analysis and classification, specifically in the context of Arabic dialects. For instance, Al-Azani and El-Alfy [11] have investigated the effect of emoji features on Arabic sentiment analysis and using the Arabic dialectical tweets. Studies have been shown that combining emojis with word embeddings and traditional models, such as Bag-of-Words and Latent semantic analysis, leads to noticeable improvements in Arabic sentiment analysis. Also, CLASENTI framework [9] was also investigated and offered a valuable approach to Arabic sentiment analysis by categorizing sentiments for more nuanced language processing. This framework highlights the study of creating rich datasets and dictionaries and their final impact on Arabic sentiment classification. Hamdi *et al*. [9] showed that combining corpus-based and lexicon-based methods improved the overall classification results of Arabic sentiment. Moreover, the Aspect-Based Sentiment Analysis (ABSA) approach was also explored in the context of Arabic sentiment analysis using various machine learning methods. For example, Zouidine and Khalil [10], the researchers evaluated how well Support Vector Machines (SVM) and Recurrent Neural Networks (RNNs) performed on Arabic hotel reviews. Their findings showed that the results of SVM classifier outperformed the RNNs in detecting sentiments related to specific aspects, highlighting the effectiveness of supervised machine learning approaches in Arabic ABSA tasks.

To further improve Arabic sentiment analysis, recent studies have proposed using various deep learning [17, 18], Large Language Models (LLMs) [10], and transformer models [16]. In particular, Zouidine and Khalil [10] have explored the application of LLMs for the purposes of both Arabic sentiment analysis and machine translation, emphasizing their adaptability and effectiveness in processing and analyzing complex linguistic features. A hybrid deep learning model called ArabBERT-LSTM, developed in Ref. [16], integrates transformer-based architectures with LSTM networks. It significantly improved the overall sentiment classification results, reaching a remarkable accuracy rate of 97%. Also, in Ref. [17], Almuhaya *et al*. have conducted a comparative study of various machine learning techniques on unbalanced Arabic social media data, and evaluated how well sampling techniques address the problems caused by unbalanced data in the context of Arabic sentiment analysis. Moreover, Radman and Duwairi [18] have proposed a robust deep learning approach tailored for Arabic sentiment analysis and demonstrated its superior performance across various datasets.

In addition to prior attempts, there is a growing recognition that Arabic dialects, including Palestinian Arabic, belong to the broader category of low-resource languages, which encounter distinct challenges due to the limited availability of labeled datasets, limited availability of pre-trained models, and significant linguistic variability. As highlighted in Ref. [8], recent research has emphasized building dialect-specific resources and models to address these gaps. Studies focusing on Egyptian, Gulf, and Levantine dialects have demonstrated that sentiment expressions are highly contextual and morphologically diverse, requiring customized linguistic tools and domain-specific lexicons. Moreover, low-resource NLP methods, including transfer learning, data augmentation, and cross-lingual embedding techniques, have been proposed to bridge the limited data problem in dialectal Arabic processing. These techniques enable the reuse of knowledge from high-resource

variants like MSA or English, to enhance performance on dialectal Arabic tasks.

By incorporating these insights, our work contributes to this growing body of research by building a manually annotated dataset and a sentiment lexicon specifically for Palestinian Arabic, which is an underrepresented dialect. We then validate its utility through both machine learning and lexicon-based approaches. This positions our study within the broader context of advancing sentiment analysis for low-resource and dialect-rich language settings.

Despite significant progress in modern Arabic sentiment analysis, research focusing on Arabic slang, particularly the Palestinian slang, remains limited and underdeveloped. The application of neural networks to sentiment analysis of Arabic slang is also still limited and needs further investigations. On the other hand, although there have been some efforts to create sentiment lexicons, particularly for Arabic slang, and thus develop a hybrid model that combines lexicon-based methods with machine learning, these attempts require further development to fully address the complexities of dialect-specific sentiment analysis. This highlights a valuable direction for the current study, which aims to advance sentiment analysis for colloquial Arabic by comparing machine learning and lexicon-based approaches specifically tailored to the Palestinian dialect, while also addressing challenges such as data imbalance and dialectal variability.

III. Materials and Methods

This section provides an overview of the dataset collected and used followed by an explanation of the proposed model. Typically, developing a sentiment analysis system includes several main phases, as shown in Fig. 1, including data collection, data preprocessing, lexicon creation and augmentation, sentiment classification model (machine learning-based model, lexicon-based model, and a neural network-based model), where the outcome of this system is to analyze the sentiments colloquial Arabic sentences into positive and negative.

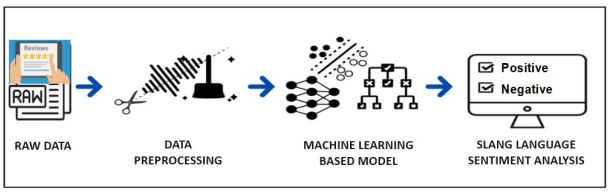

Fig. 1. The proposed workflow of a machine learning-based model for slang language sentiment analysis.

A. Data Collection and Pre-Processing

To obtain the essential dataset for training our sentiment analysis model, we started collecting customer reviews from Facebook for the main Palestinian telecom provider, Paltel. These reviews covered the services and products of prominent Palestinian telecom companies, namely JAWWAL (a mobile service provider), Paltel (a fixed line service provider), and Hadara (an internet service provider). The substantial volume of Facebook reviews on Palestinian telecom services not only provides a comprehensive sample of Palestinian slang, but also represents a diverse range of colloquial expressions within the digital landscape. This extensive dataset, consisting of 8895 reviews, was used for both training and testing the sentiment analysis model, ensuring the robustness and reliability of this research.

The dataset used in this study comprises 8895 Facebook reviews, collected from the official pages of three major Palestinian telecom providers: JAWWAL, Paltel, and Hadara. Each review was manually annotated by native Palestinian Arabic speakers as either positive or negative. The final distribution includes 6716 positive reviews and 2179 negative reviews. Neutral reviews were excluded, as this work focuses on binary sentiment classification. No data augmentation or resampling methods were applied. The dataset was deliberately maintained in its naturally imbalanced form to reflect real-world user sentiment trends observed on social media. While this class imbalance may impact certain performance metrics, particularly recall for the minority class, it provides a more realistic assessment of model robustness in real-world conditions.

To facilitate the extraction of Facebook data, we developed a novel web-based tool based on the Spring framework. This tool is adept at interfacing with the Facebook API, which enables the retrieval of data from the Facebook social network. Our tool, built on the foundation of the "http-Client" framework, orchestrates the seamless retrieval of data from Facebook. The raw data, extracted from the aforementioned sources, was inherently unprocessed, retaining its original state. This dataset underwent a series of preparatory steps to transform it into a structured and accessible format for sentiment analysis. In particular, these preparatory steps were performed as follows:

- **Transformation into a structured format:** The raw dataset, originally in SQL, was transformed into an Excel format, to facilitate subsequent processing steps.
- **Removal of irrelevant fields:** Extraneous fields not relevant to the sentiment analysis endeavor were systematically removed from the dataset, streamlining it for subsequent analysis.
- **Manual annotation of labels:** The proposed sentiment analysis involved a team effort to label sentiments. Multiple expert annotators, working in sync, assigned positive or negative labels to each data instance. Through ongoing discussions, we sought a common understanding and ensured consistency as measured by inter-annotator agreement. This collaborative approach added a nuanced human touch to the sentiment labels, enriching the dataset.

The culmination of these preparatory steps resulted in a comprehensive labelled dataset. The dataset contains a total of 8895 instances, distributed between positive and negative comments: with 6716 positive, and 2179

negative comments as shown in Table I. This table also shows some statistics including the total number of words, the average number of words, and the maximum and minimum number of words for each sentence. It can be seen that the labelled data is unbalanced, where the majority of the reviews are marked as positive comments, which could affect the classification results. Furthermore, comprehensive statistical insights into the comment structure were generated, providing a nuanced perspective on the linguistic attributes of the dataset, see Table I for more details.

TABLE I. SOME STATISTICS ABOUT THE EXTRACTED FACEBOOK REVIEWS AND HOW THEY ARE DISTRIBUTED BETWEEN POSITIVE AND NEGATIVE COMMENTS

Data Category	Total Instances	Positive Comments	Negative Comments
Count	8895	6716	2179
Total words	62,265	39,624	26,365
Average words	7.0	5.9 words	12.1 words
Maximum words	597	671	264
Minimum words	1	1	1

The computed linguistic attributes, such as total words, average words, maximum words, and minimum words, provide a comprehensive view of the comment structure, which is crucial for subsequent sentiment analysis. A sample of the gathered data is shown in Table II.

TABLE II. SAMPLE FROM THE COLLECTED AND LABELLED COMMENTS FOR JAWWAL COMPANY

Comment in Arabic	Comment in English	Label
جوال كل يوم جديد كل الاحترام	With all due respect, JAWWAL is new every day.	Positive
موفق والى الأمام	Good luck and keep moving forward	Positive
ايش يعني	what does it mean	Negative
وين حملات الفواتير؟	Where are the billing campaigns?	Negative
كل الحب لكم ولمصداقيتكم	Much love to you and your credibility	Positive
الله يوفقكم يا شركة جوال الغالية	May God grant you success, dear JAWWAL Company.	Positive
الحمد لله لم يصان شيء	Thanks God nothing was preserved	Negative

The sentiment lexicon used in this study was both created and adapted to capture the unique characteristics of Palestinian Colloquial Arabic. It comprises 7027 sentiment-bearing words and expressions. We adopted a hybrid construction method combining manual curation and algorithmic extraction. Two native Palestinian speakers manually contributed slang and dialect-specific terms that are often absent from formal Arabic lexicons. Simultaneously, we performed frequency analysis on the dataset to identify commonly occurring words in emotionally charged comments. These candidates were then validated by linguistic experts. This two-stage process ensured that the lexicon accurately reflects real-world usage in Palestinian social media and provides comprehensive coverage of sentiment-laden expressions.

All comments were preprocessed by applying the stemming, tokenization, and removal of stop words, and we represented the preprocessed comments using the Term Frequency–Inverse Document Frequency (TF-IDF) vectorization technique, which converts text into numerical features. Then we divided the dataset into two groups; 80% for training the model and 20% of the dataset is used to test and evaluate the model performance.

B. Creation and Augmentation of Sentiment Lexicon

Our methodology focuses on building a comprehensive sentiment lexicon tailored to Palestinian Colloquial Arabic. A lexicon, in the context of language processing and sentiment analysis, refers to a collection of words with their emotional or sentimental values. It is not only a matter of compiling words, but also of adding terms and phrases used in social media slang. Linguistic modifiers such as negations and intensifiers are also considered to increase the accuracy of informal Arabic sentiment analysis. Also, we take a close look at linguistic elements, including the complexity of Arabic sentences and slang expressions, that influence sentiment. For example, a word like “happy” is positive, and adding a negation like “not happy” can make that word negative or neutral. Even in negative statements, certain negations change the implied meaning of a word. These include intensifiers and diminishers, which increase or decrease the sentiment. On the other hand, the sentence “مش or مو حلو”, which means it is not nice or something bad, is typically used to denote and express bad feelings about a certain matter. As shown in Table III, our approach also covers other linguistic issues such as modals, intensifiers and many others. This specialized lexicon covers the special features of Palestinian Colloquial Arabic. It has been created through manual and automated processes, to ensure that it reflects the nuances of the language. This lexicon can improve the accuracy of sentiment analysis for Arabic social media by incorporating both linguistic nuances and cultural insights.

TABLE III. ARABIC LANGUAGE CONTROLS

Controls	Meaning in Arabic	Example in Arabic	Action
Negators	أدوات النفي	لا، لم، لا	Manipulate
Modals	الأفعال الناقصة	أصبح، أمسى	Remove
Intensifiers	التوكيد	تكرار الكلمة	Remove
Diminishers	المتناقضات	حب - كره	Phrases
Prepositions	حروف الجر	من، الى	Remove
Conjunctions	حروف العطف	و، او ، ثم	Remove
Connected Pronouns	ضمائر متصلة	الميم في (أنفسهم)	Remove
Demonstrative Pronouns	ضمائر الاشارة	هذا، هذه، هؤلاء	Remove
Relative Pronouns	الأسماء الموصولة	الذي، التي، الذين	Remove
Pronouns	الضمائر	هو، هي	Remove
Paronomasia	التورية	أعيني جودا	Phrases
Insinuation	التلميح	لا والله بطل!	Phrases

The lexicon was independently annotated by two native Palestinian Arabic speakers, and a third linguistic expert was consulted to resolve any disagreements. Most discrepancies emerged from context-sensitive terms or expressions that could convey contrasting sentiments depending on usage. For example, the word “مولع” may express excitement (positive) or frustration (negative), depending on the situation. Similarly, “يا سلام” can signal genuine admiration or be used sarcastically to indicate

disapproval. These cases were carefully reviewed in context to determine the most representative sentiment label. This validation process ensured consistency and helped refine the lexicon to better capture the nuanced emotional expressions typical of Palestinian colloquial Arabic.

The contribution to the lexicon study is the construction of a specialized Arabic Sentiment Lexicon specific to Arabic Facebook comments on social and public issues, with a focus on Palestinian slang (or dialect). The lexicon contains 7027 terms. This included incorporating sentiment lexicons translated from English into Arabic, followed by careful revision to align with the study goals. About 88% of these translations effectively captured Arabic sentiment expressions, while about 10% suffered from translation challenges [20]. In addition, we integrated an external lexicon containing 1367 words and their importance weights, which was obtained from the National Research Council in Canada. This lexicon provides English and Arabic sentiment terms, as well as sentiment annotated corpora, increasing the depth and coverage of our lexicon resources [21]. An example of the lexicon used is shown in Table IV.

TABLE IV. SAMPLE FROM THE COLLECTED AND LABELLED COMMENTS FOR JAWWAL COMPANY

Lexicon in Arabic	Lexicon in English	Weight
مبروك	Congrats	0.963
فخر	Honor	0.95
مبدع	Creative	0.938
نجاح	Success	0.912
سرور	Pleasure	0.912
تحفيز	Motivation	0.9
صادق	Honest	0.9
متفائل	Optimistic	0.9
سعيد	Happy	0.887

All models used in this study, including the MLP, were trained entirely from scratch without the use of any pre-trained embeddings or fine-tuning techniques. The MLP was configured with the following parameters: 100 training epochs, a batch size of 64, and a learning rate of 0.001. The Adam optimizer was employed during training. The model architecture consisted of a single hidden layer with 100 neurons and a ReLU activation function. The output layer used a sigmoid activation function, and binary cross-entropy served as the loss function. Although we initially adopted an 80/20 train-test split, we further implemented 5-fold cross-validation to ensure robustness in our evaluation and minimize the impact of data partitioning on model performance.

The culmination of manual and automated processes, along with our awareness of contextual and linguistic nuances, led to the creation of a sentiment lexicon tailored to Palestinian Colloquial Arabic, consisting of 7027 terms. This lexicon provides a solid foundation for accurate sentiment analysis within the realm of colloquial expressions, enhanced by the integration of linguistic subtleties, cultural insights, and external lexicon resources. This lexicon is used to build a lexicon-based sentiment model for the views and comments extracted from social media.

C. Sentiment Classification Model

The proposed sentiment analysis model involved a well-structured approach designed to evaluate model effectiveness and assess the impact of our built sentiment lexicon. The sentiment model building process consisted of three stages, namely machine learning classification, lexicon-based sentiment classification, and the combined classification model. The performance of the three models is compared with each other to see how they perform in the context of Palestinian Colloquial Arabic sentiment.

1) Machine learning classification model

In this phase, we focused on training and testing machine learning classification algorithms using the labeled dataset. We used the following classifiers: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and SVM. In particular, the LR is a statistical technique, typically used for binary classification tasks rather than regression tasks. This technique models the probability of a given input belonging to a particular class using a logistic function. Due to its simplicity and interpretability, it has often served as the basic machine learning algorithm for the classification problems. DT is another interpretable machine learning algorithm that has been used for both classification and regression tasks. The DT technique is a tree-based supervised machine learning algorithm, where each path from a root node to a leaf node is represented as a sequence of data separations until a classification result is determined at the leaf node. The splitting process is primarily based on the concept of information gain, which quantifies the amount of insights gained from a variable within the dataset. The RF is also a supervised classifier that can be used for both regression and classification analysis. The basic idea behind this model is to build a set of DTs from the training data and make predictions by aggregating the results of multiple trees. RF has been shown to provide a high classification rate and robustness to outliers and noise, thus mitigating the risk of overfitting. NB is a probabilistic classification technique, which is based on Bayes' theorem with the "Naïve" assumption, which ensures independence between features. Despite its simplicity and computational efficiency, where it requires a small amount of training data to estimate the model parameters, NB often performs well, especially on text classification tasks. In addition, KNN is a simple algorithm used for classification tasks. It classifies data points based on the majority class among their K nearest neighbors. In addition, SVM is a powerful supervised learning algorithm used for classification, regression, and outlier detection. It finds the optimal hyperplane that separates data points of different classes with the maximum margin. SVMs are effective in text analysis due to their ability to handle high-dimensional data and identify complex decision boundaries.

Although the above listed classifiers are well-suited for sentiment analysis tasks due to their ability to model the relationships between the textual features and sentiment labels. Other recent techniques have emerged, such as MLP: a type of Feedforward Neural Networks (FFNNs). MLP offers some advantages over classical machine learning algorithms, such as its ability to model complex relationships, treat high dimensional data such as colloquial Arabic, in addition to generalize the model to invisible data. MLP consists of multiple layers of nodes, including input layer, which receives the input data and extract the features of the text. As for the hidden layers, which lie between the input and output layer, we implemented one hidden layer with 100 neurons. Clearly it was sufficient for this classification task. The hidden layer employed the (Relu) activation function to introduce non-linearity into the model. The output layer of the model consists of a single node and is responsible for generating the output for the sentiment classifier (positive or negative). The MLP model is trained using the labeled dataset of the reviews text, and with the help of "Adam" optimizer, the model's weights were improved through adjusting the learning rates based on the first and second moments of the gradients.

2) Lexicon-based sentiment classification model

In this phase, the sentiment is determined by referring to the predefined lexicon and its associated sentiment scores, that we have already constructed. In this method, each word in the text is compared with all the entries in the lexicon, and its sentiment score is assigned based on the predefined values in the lexicon. The sentiment level is then estimated as the overall average of all sentiment scores collected for the text.

IV. Results

To evaluate the performance of our sentiment analysis framework, we used a dataset consisting of 8895 comments, categorized into 6716 positive and 2179 negative sentiments. To ensure an accurate evaluation, we used N-fold cross-validation and chose 5-fold cross-validation to accommodate for initial computational limitations. We employed the evaluation methods provided by SKlearn to assess the effectiveness of the algorithms used to perform the sentiment classification process. Accuracy, precision, recall and F1-Score were some of the metrics utilized to measure the performance.

The first part of the study used the same methodology followed by most of the published research, see for example [10, 22–24] where several machine learning techniques were investigated for various datasets and promising results were obtained. The second part of the experiments conducted focused on the use of the lexicon-based model, which has already been investigated in the literature, see for example [25]. However, we have used these methodologies to analyze and classify the Arabic slang language, which has not yet been considered. Furthermore, the third part of the experiments is based on using NNs for the slang language.

The results section presents the outcomes of a comprehensive analysis of the proposed sentiment analysis approaches for Arabic slang text, including the lexicon-based approach. This study aimed to discuss the feasibility of overcoming the challenges of sentiment analysis of Arabic colloquial content, which is characterized by its informality and high degree of randomness. By employing a specialized lexicon and a given scoring mechanism, we attempted to capture the sentiment nuances inherent in such texts. The performance of the lexicon-based sentiment model is compared to models built using machine learning classification algorithms.

The machine learning models were trained and optimized using Scikit-Learn in Python. The models were trained on the labelled dataset using several algorithms including LR, DT, RF, NB, KNN, and SVM.

Grid search with cross-validation using "GridSearchCV" method from Scikit-learn was used to tune and optimize the model hyperparameters or each classification algorithm. GridSearchCV is designed to search over hyperparameters and to try different combinations of values for these hyperparameters to determine the best combination to be used in the model. The same is true for MLP classifiers, where GridSearchCV iterates over all combinations of hyperparameters such as hidden layer size, activation function, solver, and others. The best combination of hyperparameters is adopted based on the best accuracy that the model can achieve.

Table V provides a comprehensive overview of the results obtained from our sentiment analysis process. The table is divided into different approaches, each of which uses specific classification algorithms and performance measures.

TABLE V. The Comprehensive Results of the Sentiment Analysis

Approach	Classification Algorithm	Performance Measure (Percentage)			
		Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Machine Learning-Based Approach	Logistic Regression	93.7	93.7	93.7	93.7
	Decision Tree	92.2	92.2	92.2	92.2
	Random Forest	93.9	93.9	93.9	93.9
	Support Vector Machine	93.9	93.9	93.9	93.9
	K-nearest neighbors	81.0	81.0	81.0	81.0
	Naïve Bays	94.0	94.0	94.0	94.0
Neural Network-Based Approach	MLP Classifier	95.0	95.0	95.0	95.0
Lexicon-Based Approach	-	81.0	77.0	64.0	67.0

These results show that, except for KNN, all classification algorithms showed remarkably high performance, achieving accuracy rates ranging from 92.2% to 94% with only a slight 2% variation. Naïve Bayes stood out as the best performer in terms of performance measures, while KNN showed the lowest performance. The relatively high accuracy of the different algorithms indicates that they can serve as effective tools in sentiment analysis of Arabic Slang language, and that multiple algorithms can definitely achieve comparable results.

When using the MLP, a type of FFNN classifier as previously mentioned, it achieves slightly better performance than classical machine learning algorithms, with an accuracy of 95% as shown in Table V. However, this advantage is modest due to the relatively small size of the dataset, which limits the neural network's ability to fully leverage its strengths. Increasing the dataset size is expected to significantly improve the performance of MLP's classification, further distinguishing it from other traditional methods.

As for using the prepared lexicon in predicting the sentiment label of the data used, the results showed a lower performance than the machine learning algorithms; with an accuracy of (81%), precision of (77%), recall of (64%) and F-score of (67%). The lower performance of the lexicon-based approach could be attributed to the ineffectiveness of the prepared lexicon and its inappropriateness for the specific dialect or the language variant being analyzed. Moreover, the unique expressions, nuances, and emotional cues of the Palestinian colloquial dialect may not have been professionally addressed by the dictionary because it may not have been specifically designed for this dialect. In addition, lexicon-based approaches can be limited by the coverage and quality of the lexicon, which may affect their performance. Overall, while the lexicon-based approach showed promising results, further refinement and expansion of the lexicon, as well as possibly combining it with machine learning techniques, could help improve its performance in sentiment analysis of Arabic slang language.

The observed gap between the accuracy (81%) and the F1-Score (67%) of the lexicon-based model is primarily due to the imbalance between classes in the dataset. As the majority of the reviews are positive, the rule-based model tends to favor the dominant class. This results in higher accuracy but poor recall for the minority (negative) class, thereby lowering the F1-Score. Moreover, the lexicon-based approach lacks the contextual adaptability of learning-based models relies on fixed sentiment scores. This limitation hinders its ability to accurately interpret nuanced or sarcastic negative expressions. In contrast, machine learning models such as MLP and Naïve Bayes demonstrated better handling of class imbalance and delivered more balanced performance across both sentiment classes.

Although the dataset imbalance was previously acknowledged, it is important to highlight how this skew can influence evaluation metrics such as accuracy and F1-Score. Due to the predominance of positive reviews, models may achieve high scores by favoring the majority class, which may lead to missing negative sentiments. Although resampling techniques were not applied in this study, we recognize their value and intend to investigate methods such as oversampling and under-sampling in future work. These methods will allow us to assess whether balancing the distribution of classes enhances performance, particularly for the minority class.

V. Discussion

The sentiment analysis results show differences between the ML algorithms and the lexicon-based technique. It is worth noting that the ML algorithms consistently outperform the lexicon-based approach across all performance metrics. This difference can be attributed to the ML algorithms inherent ability to understand relationships between words and sentiment while adapting to nuances in the data. The lexicon-based technique; on the other hand, relies on predefined sentiment scores or weights for words and lacks the adaptive nature of ML algorithms. In addition, it is important to note that the precision scores, for machine learning algorithms are significantly higher indicating their ability to classify sentiments as positives. This is because these algorithms can learn from labelled data and generalize sentiment patterns beyond what is predefined in a lexicon. The same observations apply to the Recall and F1-Score indicators.

The lexicon-based approach achieved an accuracy of 81%, demonstrating its moderate effectiveness in sentiment analysis for the Colloquial Arabic language. The differences between ML algorithms and lexicon-based approaches lie in the methods they use. Machine learning algorithms rely on sets of labelled training data to understand patterns and relationships between words and sentiments. This allows them to make sentiment predictions that go beyond what the lexicon approach can do. On the other hand, the lexicon-based approach relies on predefined sentiment scores for words, which means it lacks the adaptability and context awareness that machine learning algorithms naturally possess. Furthermore, the lexicon-based approach can struggle to capture nuances, sarcasm, or language-specific to domains. These are areas where machine learning models excel because of their data-driven learning processes.

While the lexicon-based model performed less well than expected compared to machine learning methods, this cannot be attributed solely to limitations in lexicon size or word coverage. A more critical analysis reveals that sentiment expression in Palestinian dialect, especially in informal social media contexts, is highly context-dependent. Many dialectal terms exhibit sentiment polarity that changes based on sentence structure, tone, or surrounding words. Furthermore, the lexicon-based approach struggles to accurately process linguistic modality such as negations (e.g., "not good"), sarcasm, and intensifiers (e.g., "very bad" vs. "bad"), which are prevalent in user-generated content. These

challenges undermine the reliability of fixed rule-based scoring. Our findings suggest that a hybrid approach, combining the linguistic insights of sentiment lexicons with the contextual learning capabilities of machine learning, could offer more robust sentiment detection. We intend to pursue this integration in future work to improve performance across informal and dialect-rich datasets.

It can be seen that the results of the MLP outperform the results of the classical machine learning algorithms. This is due to the fact that MLP classifiers can model complex non-linear relationships between the input features and the output labels. They can also automatically extract relevant features from the input data during the learning phase, where the traditional machine learning techniques require manual feature engineering. This can be time-consuming and does not guarantee capturing all relevant features in the data. In addition to that, MLP classifiers are able to learn hierarchical representations of data through multiple hidden layers of nodes. This allows them to capture high-level abstractions and complex patterns in the data, leading to improved performance on challenging tasks. These promising results encourage the authors to conduct further investigations, especially combining MLP with lexicon-based models.

VI. Conclusions and Future Works

This study explored sentiment analysis of Palestinian Colloquial Arabic by comparing lexicon-based methods, traditional machine learning algorithms, and a MLP classifier. The analysis highlights the limitations of applying standard formal Arabic tools to dialectal content, particularly in social media, where slang and informal expressions are prevalent. Among the evaluated methods, the MLP classifier achieved the highest performance, outperforming both the classical machine learning techniques and the lexicon-based model. This result aligns with existing literature demonstrating the capacity of neural networks to capture complex language patterns and subtle semantic variations, particularly in low-resource languages and dialects.

The relatively poor performance of the lexicon-based approach underscores the challenges of relying on static word lists that may not reflect the fluid and expressive nature of colloquial Arabic. This highlights the need for more solid and context-aware lexicons tailored to dialectal data. The findings suggest that while machine learning and neural models provide better adaptability and generalization, hybrid approaches combining deep learning with domain-specific lexicons may further enhance sentiment classification accuracy.

Future research efforts should focus on improving domain lexicons designed specifically for slang text. Important research topics include studying the differences in sentiment expression between dialects and diving into deep learning models. Another approach that aims to improve sentiment analysis in slang language by providing more accurate and contextual results is sentiment intensity prediction. In addition, increasing the size of the dataset would improve the classification results, using the MLP model in particular, which will pave the way for proposing a combined deep learning model [26], which will be based on deep neural networks and a lexicon-based model.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

KRan prepared the datasets, conducted the experiments, and developed the models; KR and AH analyzed and investigated the results, and together with KRan, contributed to writing the paper; all authors had approved the final version.

References

[1] A. Silberling. (July 2023). Facebook surpasses 3 billion monthly active users. *TechCrunch*. [Online]. Available: https://techcrunch.com/2023/07/26/facebook-3-billion-users/

[2] G. Alwakid, T. Osman, and T. Hughes-Roberts, "Challenges in sentiment analysis for Arabic social networks," *Procedia Computer Science*, vol. 117, pp. 89–100, 2017.

[3] A. Ghallab, A. Mohsen, and Y. Ali, "Arabic sentiment analysis: A systematic literature review," *Applied Computational Intelligence and Soft Computing*, vol. 2020, pp. 1–21, 2020.

[4] N. Boudad, R. Faizi, R. O. H. Thami *et al.*, "Sentiment analysis in Arabic: A review of the literature," *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 2479–2490, 2018.

[5] Y. Matrane, F. Benabbou, and N. Sael, "A systematic literature review of Arabic dialect sentiment analysis," *J. King Saud Univ.–Computer and Information Sciences*, vol. 35, no. 6, 101570, 2023.

[6] R. Obiedat, D. Al-Darras, E. Alzaghoul *et al.*, "Arabic aspect-based sentiment analysis: A systematic literature review," *IEEE Access*, vol. 9, pp. 152628–152645, 2021.

[7] S. Tartir and I. Abdul-Nabi, "Semantic sentiment analysis in Arabic social media," *J. King Saud Univ.–Computer and Information Sciences*, vol. 29, no. 2, pp. 229–233, 2017.

[8] S. M. Sherif, A. H. Alamoodi, O. S. Albahri *et al*., "Lexicon annotation in sentiment analysis for dialectal Arabic: Systematic review of current trends and future directions," *Information Processing & Management*, vol. 60, no. 5, 103449, 2023.

[9] A. Hamdi, K. Shaban, and A. Zainal, "Clasenti: A class-specific sentiment analysis framework," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 4, pp. 1–28, 2018.

[10] M. Zouidine and M. Khalil, "Large language models for Arabic sentiment analysis and machine translation," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20737–20742, 2025.

[11] S. Al-Azani and E. S. M. El-Alfy, "Combining emojis with Arabic textual features for sentiment classification," in *Proc. 9th Int. Conf. Inf. Commun. Syst. (ICICS)*, 2018, pp. 139–144.

[12] N. A. Abdulla, N. A. Ahmed, M. A. Shehab *et al*., "Towards improving the lexicon-based approach for Arabic sentiment analysis," *Int. J. Inf. Technol. Web Eng.*, vol. 9, no. 3, pp. 55–71, 2014.

[13] Y. Alqahtani, N. Al-Twairesh, and A. Alsanad, "Improving sentiment domain adaptation for Arabic using an unsupervised self-labeling framework," *Information Processing & Management*, vol. 60, no. 3, 103338, 2023.

[14] A. Mountassir, H. Benbrahim, and I. Berrada, "Sentiment classification on Arabic corpora: Preliminary results of a cross-study," in *Proc. 3e Séminaire de Veille Stratégique, Scientifique et Technologique (VSST)*, 2012, 12.

[15] M. Al-Ayyoub, S. B. Essa, and I. Alsmadi, "Lexicon-based sentiment analysis of Arabic tweets," *Int. J. Soc. Netw. Mining*, vol. 2, no. 2, pp. 101–114, 2015.

[16] W. Alosaimi, H. Saleh, A. A. Hamzah *et al*., "ArabBert-LSTM: Improving Arabic sentiment analysis based on transformer model

and long short-term memory," *Frontiers in Artificial Intelligence*, vol. 7, 1408845, 2024.
[17] B. Almuhaya, B. Saha, M. Kaur *et al*., "Comparative analysis of machine learning algorithms for Arabic sentiment analysis on imbalanced social media data," in *Proc. ASU Int. Conf. Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, 2024, pp. 1362–1367.
[18] A. Radman and R. Duwairi, "Towards a robust deep learning framework for Arabic sentiment analysis," *Natural Language Processing*, vol. 31, no. 2, pp. 500–534, 2025.
[19] H. Alawneh, A. Hasasneh, and M. Maree, "On the utilization of emoji encoding and data preprocessing with a combined CNN-LSTM framework for Arabic sentiment analysis," *Modelling*, vol. 5, no. 4, 2024.
[20] M. Salameh, S. Mohammad, and S. Kiritchenko, "Sentiment after translation: A case-study on Arabic social media posts," in *Proc. the 2015 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 767–777.
[21] M. Salameh, S. M. Mohammad, and S. Kiritchenko. (2016). Arabic sentiment analysis and cross-lingual sentiment resources. [Online]. Available: http://saifmohammad.com/WebPages/ArabicSA.html
[22] O. Araque, I. Corcuera, C. Román *et al*., "Aspect based sentiment analysis of Spanish tweets," in *Proc. TASS 2015: Workshop on Sentiment Analysis at SEPLN Co-located with 31st SEPLN Conference (SEPLN 2015)*, 2015, pp. 29–34.
[23] S. Mukherjee and P. Bhattacharyya, "Sentiment analysis in Twitter with lightweight discourse analysis," in *Proc. COLING 2012: Technical Papers*, 2012, pp. 1847–1864.
[24] T. H. Soliman, M. A. Elmasry, A. Hedar *et al*., "Sentiment analysis of Arabic slang comments on Facebook," *Int. J. Comput. Technol.*, vol. 12, no. 5, pp. 3470–3478, 2014.
[25] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis—A hybrid scheme," *Journal of Information Science*, vol. 42, no. 6, pp. 782–797, 2016.
[26] Y. Zayed, A. Hasasneh, and C. Tadj, "Infant cry signal diagnostic system using deep learning and fused features," *Diagnostics*, vol. 13, no. 12, 2107, 2023.