International Journal of Electronic Commerce Studies
Vol.16, No.3, pp.23-48, 2025
doi: 10.7903/ijecs.2528

Sentiment Analysis for E-commerce in the Maghreb: Enhancing Algerian Dialects Classification with BERT

Faiz MAAZOUZI
Mohamed Cherif Messaadia University, Algeria
f.maazouzi@univ-soukahras.dz

Ahmed AHMIM
Mohamed Cherif Messaadia University, Algeria
a.ahmim@univ-soukahras.dz

Massifa Messadia
Mohamed Cherif Messaadia University, Algeria
messaadiamassifa@gmail.com

ABSTRACT

E-commerce platforms have become essential in meeting diverse consumer needs rapidly. For instance, Jumia—the largest e-commerce platform in North Africa—receives a high volume of user reviews that reflect a wide range of opinions regarding products. This diversity challenges platform owners striving to offer high-quality products and leaves buyers uncertain about making the best choices. To address these issues, we developed a sentiment analysis framework specifically tailored to the Algerian dialect. Our approach involved constructing a comprehensive database of user reviews categorized into positive, negative, and neutral sentiments. We further enhanced this resource by compiling a specialized dictionary of commonly used Algerian terms and applying GAN-based expansion techniques, as well as translating reviews into English and French to broaden linguistic coverage. To evaluate our method, we implemented two deep learning classifiers: a Deep Neural Network (DNN) and a BERT-based model. Notably, the BERT model achieved its optimal performance at 20 training epochs, with an accuracy of 95.44%, precision of 93.1%, recall of 95.57%, and an F1-score of 94.7%. These results significantly surpassed those obtained using the DNN model, as confirmed by ROC curve analyses and comparative accuracy evaluations. Our findings demonstrate that the integration of advanced NLP techniques with domain-specific language resources markedly enhances sentiment classification, paving the way for more effective analysis systems in e-commerce applications and the broader incorporation of Maghrebi dialects into scientific research.

Keywords: Sentiment Analysis, Maghreb Dialect, Data Augmentation, Lexicon-Based Approach, classification, BERT

1. INTRODUCTION

The Internet has become an immensely popular and cost-effective medium for sharing information, particularly through social media. Platforms such as blogs, reviews, posts, and tweets are analyzed to extract public opinions on various topics, a process known as sentiment analysis or opinion mining. This analysis enables organizations to understand and categorize sentiments, helping them respond more effectively. The success of sentiment analysis depends on specific objectives, such as identifying text polarity, sentiment, distinctive reactions, or even language detection.

For our study, we selected Jumia, the leading e-commerce platform in the Maghreb region, as our focus. Given its widespread popularity, classifying customer opinions on its products presents a significant challenge. This challenge is further compounded by the need to incorporate sentiment analysis for the Algerian dialect, which is inherently complex and rich due to the diverse civilizations and cultures that have influenced it. As a result, it is essential to employ advanced techniques and algorithms capable of accurately detecting emotional trends and audience reactions, even within the intricate linguistic context of North African dialects.

By analyzing natural language and identifying key indicators of positivity, negativity, or neutrality, these systems provide valuable insights into people's attitudes and emotions. This allows for a deeper understanding of the opinions expressed in diverse linguistic environments, such as North African dialects. In our research, we aimed to develop a sentiment analysis model with a novel architecture designed to deliver accurate assessments. To achieve this, we followed several key stages:

First, we manually collected comments in the Algerian dialect from the Jumia website and classified each one as positive, negative, or neutral. This step laid the foundation for our project. Next, we employed various methods to expand and enrich the database. This included creating a comprehensive dictionary that captures a significant portion of the Algerian vocabulary and incorporating foreign languages, such as English and French, into the comments. Additionally, we experimented with other techniques to further enhance the quality of our results.

Finally, we designed and trained a deep learning model specifically tailored to classify comments as positive, negative, or neutral. This process involved pre-processing the data, selecting the appropriate model architecture, training the model on the augmented database, and evaluating its performance. Once validated, the model is deployed to automatically classify comments on the Jumia platform.

1.1 Theoretical Framework and Contributions

Recent studies in sentiment analysis have predominantly focused on standard language forms, often overlooking the complex interplay of linguistic, cultural, and contextual factors inherent in dialects [21]. Our work builds upon established theories in computational linguistics and sociolinguistics that emphasize the significance of

linguistic variability in sentiment expression. By addressing the unique challenges of the Algerian dialect, our study advances these frameworks in several key ways:

- **Integration of Linguistic Theories with Deep Learning:** Traditional sentiment analysis methods frequently fail to capture the nuances of dialectal language. Our approach leverages advanced deep learning models—specifically, BERT and DNN—to process and analyze sentiment in the Algerian dialect. This integration not only validates theoretical constructs regarding linguistic variation and sentiment but also demonstrates that dialect-specific features can be effectively captured by modern NLP techniques.

- **Empirical Support for Dialectal Nuance in Sentiment Detection:** The superior performance of the BERT model, which achieved an accuracy of 95.44% and an F1-score of 94.7% at 20 training epochs, provides strong empirical evidence for the theoretical claim that incorporating dialect-specific data improves sentiment classification. These findings support the hypothesis that dialectal expressions contain distinct sentiment markers, thereby reinforcing theoretical perspectives on the role of cultural and linguistic diversity in computational models.

- **Expanding the Scope of Sentiment Analysis:** By applying GAN-based vocabulary expansion alongside state-of-the-art deep learning methods, our study extends current sentiment analysis models to account for low-resource dialects. This methodological innovation bridges the gap between empirical performance and theoretical understanding, offering a blueprint for incorporating underrepresented linguistic variations into broader sentiment analysis frameworks.

In summary, our contributions lie not only in the significant performance improvements demonstrated through our empirical results but also in the theoretical advancement of sentiment analysis. Our work challenges and enriches existing frameworks by providing a comprehensive model that accounts for dialectal nuances—paving the way for future research in multilingual and dialect-specific sentiment analysis.

1.2 Paper organization:

The rest of this paper is organized as follows. In Section 2, we review the existing literature in sentiment analysis with a focus on the challenges in processing the Algerian dialect and other Maghreb languages. Sections 3 and 4 details the methodology, including the data collection process, augmentation techniques, pre-processing steps, and the deep learning architectures (BERT and DNN) employed in this study. Section 5 presents the experimental results, showcasing evaluation metrics, comparative analyses, and visualizations that illustrate the performance of our models. In Section 6, we discuss the findings, address model generalizability, and outline the study's limitations. Finally, Section 6 concludes by summarizing the contributions of this work and suggesting directions for future research.

2. RELATED WORK

A review of research in sentiment analysis reveals that the majority of studies concentrate on English texts, resulting in the development of high-quality frameworks and tools. In contrast, there is a notable lag for other languages, including Arabic, particularly its dialects from the Arab world. This indicates a need for further research to create more precise tools for these languages. Recently, there has been growing interest in sentiment analysis for Arabic, especially the Maghrebi dialect, with researchers increasingly utilizing the Internet as a primary source for gathering comments in various dialects. The analysis of sentiments in Arabic dialects, particularly those found on social media platforms like Facebook, has attracted increasing attention. Various datasets and machine learning techniques have been employed to explore sentiment analysis in North African dialects, especially the Algerian and Tunisian dialects. Below is a summary of related works that highlight significant contributions in this.

2.1 Facebook as a Data Source and Machine Learning Techniques

Social media platforms, particularly Facebook, have become critical sources for collecting dialectal Arabic data, enabling researchers to refine sentiment analysis techniques through machine learning and deep learning approaches. The studies discussed below illustrate a progression in addressing the challenges of sentiment

analysis across various Arabic dialects, leveraging Facebook-derived datasets to evaluate and enhance model performance.

In [1], the Tunisian Sentiment Analysis Corpus (TSAC), compiled from Facebook, was utilized to investigate sentiment analysis in Tunisian dialects. The study assessed models such as Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), and Naive Bayes (NB), with Naive Bayes exhibiting a notable error rate of 42%. This high error rate highlights the complexity of Tunisian dialectal text, setting the stage for subsequent research to explore more robust methodologies. Building on the challenges identified in [1], the study [2] analyzed the "Wacht7ass" dataset, sourced from Algerian Facebook pages, to evaluate sentiment classification performance. Models including SVM, Naive Bayes, and Decision Trees were tested using the F-score metric, with SVM achieving scores between 76% and 87%. This improvement over the Naive Bayes performance in [1] suggests that SVM may better handle dialectal variations, particularly in Algerian contexts, prompting further exploration of model adaptability. Similarly, in [3], the DziriOFN corpus, comprising over 8,700 texts from public Facebook pages and groups, was employed to advance sentiment analysis in Algerian dialects. The study evaluated advanced models like Bidirectional Long Short-Term Memory (BiLSTM), Convolutional Neural Networks (CNN), and Naive Bayes, with Naive Bayes achieving a respectable accuracy of 75.2%. By incorporating deep learning techniques, [3] extends the findings of [2], demonstrating incremental progress in balancing accuracy and complexity for Algerian dialect processing.

Furthering this trajectory, the research [4] examined Facebook messages in both Modern Standard Arabic (MSA) and Algerian Dialect (AlgD), applying CNN and LSTM models to achieve an F1-score of 89%. This study bridges the dialectal focus of [2] and [3] with MSA, showcasing the versatility of deep learning models across linguistic variations. Collectively, these studies [1], [4] underscore a shared reliance on Facebook data to tackle the linguistic diversity of Arabic dialects, with each contributing to a deeper understanding of model efficacy and dialect-specific challenges in sentiment analysis.

2.2 Other Facebook-based Datasets and Techniques

Additional studies have leveraged data from Facebook pages to evaluate a range of models and techniques for processing Algerian dialectal content, further advancing the

application of machine learning and deep learning in this domain. These investigations, detailed below, collectively highlight the potential of Facebook as a rich source for dialect-specific analysis while demonstrating improvements in model performance and technique refinement.

In [5], data extracted from the official Facebook page of the Algerian Telephone Operator Ooredoo was utilized to assess text recognition techniques. The applied models achieved a character error rate of 10.07%, indicating a high level of accuracy in recognizing dialectal text. This study establishes a baseline for preprocessing challenges in Algerian dialect data, setting the stage for subsequent sentiment-focused analyses. Building on the text processing foundation laid by [5], the research in [6] conducted sentiment analysis on Facebook comments written in the Algerian dialect, employing Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) models. These deep learning approaches yielded an impressive F1-score of 98.09%, underscoring their exceptional capability to capture nuanced sentiments in dialectal text. The significant leap in performance from [5]'s text recognition to [6]'s sentiment analysis reflects the growing sophistication of deep learning methods tailored to Algerian dialects.

Similarly, in [7], researchers analyzed comments from Algerian Facebook pages to evaluate Multi-Layer Perceptron (MLP) and CNN models, with the CNN model achieving an accuracy of 89.5%. While this result is slightly lower than the F1-score reported in [6], it reinforces the effectiveness of CNN-based architectures across different datasets and evaluation metrics. Together, [5], [6], and [7] illustrate a continuum of progress from accurate text recognition to advanced sentiment classification using Facebook-derived Algerian dialect data, contributing valuable insights into model optimization and dialectal specificity.

2.3 Use of Datasets from Various Sources

Several studies have utilized diverse datasets for sentiment analysis across various Arabic dialects:

In [8], the AlgD dataset (ADED) was used with SVM, KNN, and LDA models, achieving an average recognition rate of 87.50%. Study [9] employed the TSAC Corpus and the Maghrebi Dialect (North Africa Corpus), applying CNN, LSTM, and BiLSTM

models. The LSTM model achieved an F1-score of 83%. In [10], the LTD corpus for the Tunisian dialect was used with a BiLSTM model, achieving an accuracy of 98.65%.

2.4 Advances in Algerian Dialect Sentiment Analysis

Several studies have advanced sentiment analysis for Algerian dialects by leveraging diverse corpora and machine learning models. For instance, the use of Time Delay Neural Networks (TDNNs) with the AlgeD corpus in [11] achieved remarkably low Word Error Rates (WER) of 1.4% for French and 7.1% for Blida and AlgeD dialects, suggesting that TDNNs are highly effective for small, annotated datasets. Similarly, the LSTM model in [12] reduced WER by 3.8% on a spoken digit corpus, indicating that recurrent neural networks excel in capturing sequential patterns in dialectal speech.

In contrast, studies like [13] and [14] reveal the complexities of dialectal variation. The 63.5% precision achieved with DNNs on the ALG-DARIDJAH corpus [13] underscores ongoing challenges with linguistic diversity and context, while [14]'s integration of French models improved Algerian dialect WER from 89% to 65.45%, hinting at the potential of cross-lingual approaches. Meanwhile, [15]'s 76% accuracy using the PADIC corpus with LSVM, BNB, and MNB models demonstrate the viability of traditional machine learning for broader Maghrebi dialects.

Beyond technical performance, handling informal language poses additional hurdles. The BERT-based approach in [16] with the NArabizi treebank, achieving an F1-score of 0.491, highlights the difficulty of mixed-language content—a common feature in informal dialects. Finally, [17]'s review from 2017 to 2024 emphasizes dataset augmentation as a critical strategy for progress, connecting the success of these models to the availability of robust, diverse data.

Together, these studies suggest that while neural networks and hybrid approaches are pushing boundaries, the interplay of dialectal variation, informal language, and data quality remains central to improving sentiment analysis for Algerian dialects.

2.5 BART and BERT Models for Sentiment Analysis

Recent studies have also explored the use of BART and BERT models for sentiment analysis in various contexts. For instance, study [18] investigated sentiment analysis of COVID-19-related tweets using the BERT model, achieving a validation accuracy of 94% on both global and Indian datasets. Similarly, in [19], the SST2 (Stanford

Sentiment Treebank) dataset was used with BERT, yielding an accuracy of 92.7%. Furthermore, study [20] focused on sentiment analysis of agricultural product reviews, where an improved BERT model achieved an F1 score of 89.86%. These findings highlight the versatility and effectiveness of BERT-based models in sentiment analysis across different domains.

3. SENTIMENT ANALYSIS WITH DEEP LEARNING

Sentiment analysis (SA) is widely used to evaluate social media users' opinions on various topics. While data mining is commonly employed, this study proposes leveraging Deep Learning to gain a deeper understanding of customer expectations and opinions [21], particularly in Big Data contexts where data mining struggles with feature identification and selection [22].

Deep Learning models iteratively learn features and generate abstract representations, making them more resilient to data variations [23]. These models are especially effective in addressing Big Data challenges, such as semantic indexing and data tagging, providing a more efficient solution for complex AI tasks. Although Deep Learning has achieved significant success in fields like computer vision, its application to Big Data sentiment analysis remains an evolving area of research [24].

4. SENTIMENT ANALYSIS PROCESS FOR THE ALGERIAN DIALECT

In this section, we describe the deep learning (DL) approach for sentiment analysis (SA) using annotations from JUMIA, specifically targeting comments written in the Algerian dialect. The process begins with collecting comments from the platform, followed by classifying each comment and applying data augmentation techniques. Subsequently, the comments undergo a cleaning and pre-processing phase. A feature selection step is also implemented to enhance the accuracy of our classification models. Finally, the evaluation phase assesses the performance of the model. Figure 1 illustrates the proposed deep learning system for sentiment analysis of Algerian comments from various regions across the country.

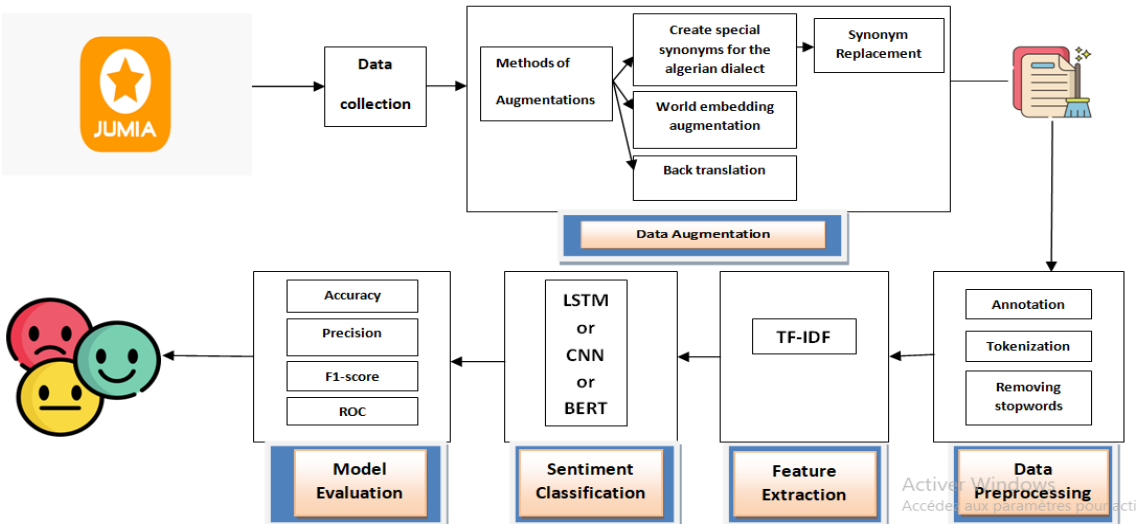


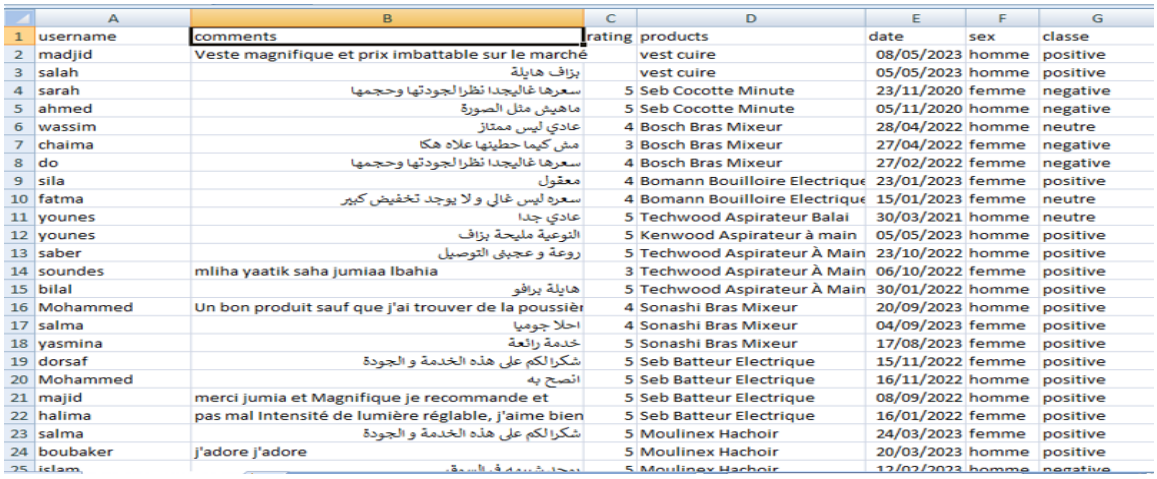
Figure 1. Proposed deep learning system for Algerian Sentiment Analysis.

4.1 DATA COLLECTION

We have developed a database based on the Jumia website, which offers a diverse range of products. Each product varies in type, purpose, and, notably, quality. As a result, this database captures all relevant details for each product, including customer names, their comments and ratings, the date of the comment, and a classification section indicating whether the comment is positive, negative, or neutral. This information is illustrated in Figure 2, while Table 1 present the type of each attribute.

Table. 1 . Data attributes and their types

Attribute	Type
used_name	String
comments	String
rating	Numeric
date	Date
name_of_product	String
sex	String
class	String



	A	B	C	D	E	F	G
1	username	comments	rating	products	date	sex	classe
2	madjid	Veste magnifique et prix imbattable sur le marché		vest cuire	08/05/2023	homme	positive
3	salah	بزاف هايلة		vest cuire	05/05/2023	homme	positive
4	sarah	سعرها غاليجدا نظرا لجودتها وحجمها	5	Seb Cocotte Minute	23/11/2020	femme	negative
5	ahmed	ماهيش مثل الصورة	5	Seb Cocotte Minute	05/11/2020	homme	negative
6	wassim	عادي ليس ممتاز	4	Bosch Bras Mixeur	28/04/2022	homme	neutre
7	chaima	مش كيما حطيتها علاه هكا	3	Bosch Bras Mixeur	27/04/2022	femme	negative
8	do	سعرها غاليجدا نظرا لجودتها وحجمها	4	Bosch Bras Mixeur	27/02/2022	femme	negative
9	sila	معقول	4	Bomann Bouilloire Electrique	23/01/2023	femme	positive
10	fatma	سعره ليس غالي و لا يوجد تخفيض كبير	4	Bomann Bouilloire Electrique	15/01/2023	femme	neutre
11	younes	عادي جدا	5	Techwood Aspirateur Balai	30/03/2021	homme	neutre
12	younes	النوعية مليحة بزاف	5	Kenwood Aspirateur à main	05/05/2023	homme	positive
13	saber	روعة و عجبني التوصيل	5	Techwood Aspirateur À Main	23/10/2022	homme	positive
14	soundes	mliha yaatik saha jumiaa lbahia	3	Techwood Aspirateur À Main	06/10/2022	femme	positive
15	bilal	هايلة برافو	5	Techwood Aspirateur À Main	30/01/2022	homme	positive
16	Mohammed	Un bon produit sauf que j'ai trouver de la poussièr	4	Sonashi Bras Mixeur	20/09/2023	homme	positive
17	salma	احلا جوميا	4	Sonashi Bras Mixeur	04/09/2023	femme	positive
18	yasmina	خدمة رائعة	5	Sonashi Bras Mixeur	17/08/2023	femme	positive
19	dorsaf	شكرا لكم على هذه الخدمة و الجودة	5	Seb Batteur Electrique	15/11/2022	femme	positive
20	Mohammed	انصح به	5	Seb Batteur Electrique	16/11/2022	homme	positive
21	majid	merci jumia et Magnifique je recommande et	5	Seb Batteur Electrique	08/09/2022	homme	positive
22	halima	pas mal Intensité de lumière réglable, j'aime bien	5	Seb Batteur Electrique	16/01/2022	femme	positive
23	salma	شكرا لكم على هذه الخدمة و الجودة	5	Moulinex Hachoir	24/03/2023	femme	positive
24	boubaker	j'adore j'adore	5	Moulinex Hachoir	20/03/2023	homme	positive
25	islam		5	Moulinex Hachoir	12/02/2023	homme	negative

Figure 1. Initial Database before augmentation

4.2 THE AUGMENTATION OF THE DATABASE

To expand our dataset, we employed the following four methods:

4.2.1 Synonym Replacement

The first method we used for data expansion was synonym replacement. We assessed whether the words belonged to our existing dialect and created a corpus incorporating various potential synonyms specific to our dialect.

4.2.2 Word Embeddings

We utilized word embeddings to identify and replace similar words or phrases within the text based on cosine similarity or other distance metrics in the embedding space.

4.2.3 Translation

Another method for enriching our dataset involved translating our Algerian dialect into the three most widely spoken languages in our country: Arabic, English, and French.

4.2.4 GAN Method

We employ Generative Adversarial Networks (GANs) to generate synthetic data that closely mimics the characteristics of real data, making them particularly useful in scenarios with limited data availability. In the context of text data, variations of GANs can be used to generate new textual content while preserving the features and patterns of the original data. Figure 3 illustrates the stages of using GANs to expand the dataset.

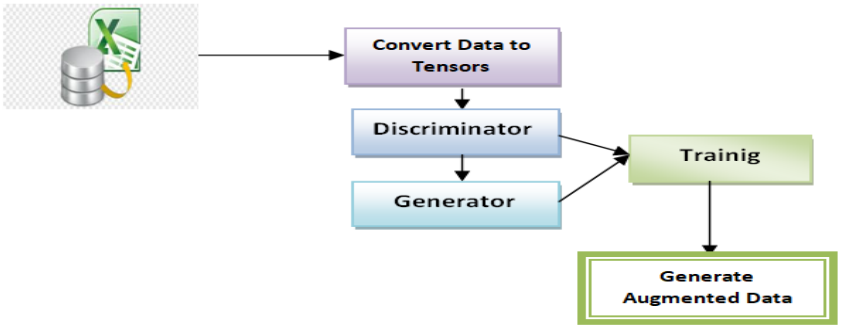


Figure 2. The processes of the Data Augmentation using GAN

After the data augmentation process, we obtained a new database containing more than 20,000 comments. Figure 4 illustrates the database after augmentation.

	A	B	C	D	E	F	G
20163	amira	bravo معجب ,الدنياش	5	Support A	########	femme	positive
20164	Messaouda	tres bien ajeb de,nia بزاف	3	Support A	########	femme	positive
20165	salma	لا mechi ,مليح	5	Support A	########	femme	negative
20166	Mohamed	bien 3 ياسر,ejbouni	5	Curren Mc	########	homme	positive
20167	samihaAZ	je suis pas satisfaite mahich, مليحا	5	Mac Style	########	femme	negative
20168	Messaouda	يعطيكم الصحة ليستها جا,تى ةbahia	5	Mac Style	########	femme	positive
20169	Baliche	khsara mechi, مليحة	5	Mac Style	########	femme	negative
20170	salma	يعطيكم الصحة ليستها جا,تى ةmleha	5	Mac Style	########	femme	positive
20171	Mohamed	non n'est pas parfait du tous mahich, mleha	5	Manteau I	########	homme	negative
20172	amira	Merci ةليستها جا,تى مليحة	5	Curren Mc	########	femme	positive
20173	Messaouda	je suis pas satisfair mechi, mleh	3	Curren Mc	########	femme	negative
20174	salma	bono ليستها جا,تى هبالة	5	Curren Mc	########	femme	positive
20175	Mohamed	yaatikm saha 3jbtn,i بزاف	5	Curren Mc	########	homme	positive
20176	samihaAZ	khsara mechi, mli7	5	Curren Mc	########	femme	negative
20177	cherif	BAHI ليستها جا,تى mli7aة	5	Curren Mc	########	homme	positive
20178	yamina	good 3 ياسر,eb denia	5	Curren Mc	########	femme	positive
20179	nourElhouda	TOP ليستها جا,تى hbalة	4	Curren Mc	########	femme	positive
20180	nourElhouda	Flop ghalia, yasser	4	Curren Mc	########	femme	negative
20181	samihaAZ	bono 3ejbet,ni bzaf	4	Support A	########	femme	positive
20182	amira	non ليس ,في المستطاع	4	Support A	########	femme	negative
20183	cherif	merci ا9it cause chi الخدمة و الاسلوب و الجودة	4	Support A	########	homme	positive
20184	yamina	bien lkaliti t,a3ha mliha	3	Support A	########	femme	positive
20185	samihaAZ	برافو ra,w3a	4	Support A	########	femme	positive
20186	naaima	TOP ichriwha, rahi top	4	Support A	########	femme	positive
20187	Messaouda	maaibetnich mana9derech nech rih khater ghali	4	Support A	########	femme	negative

Figure 3. Augmented Dataset

4.3 DATA PREPROCESSING

The pre-processing stage is a crucial step in sentiment analysis, particularly in text processing [25]. It involves several key steps, including annotation, tokenization, and the removal of stop words.

4.3.1 Annotation:

During the annotation process, we label the sentiment of each text as positive, negative, or neutral.

4.3.2 Tokenization:

Tokenization involves breaking down a text into individual words or tokens, simplifying the analysis process for algorithms. For Arabic text, we use the word_tokenize() function from Python's NLTK library, which segments Arabic text into distinct words or tokens, facilitating subsequent text processing steps [26].

4.4 Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical method used to evaluate the importance of a word within a document relative to a collection of documents (corpus). In natural language processing, TF-IDF is commonly employed to transform text data into numerical vectors, enabling machine learning models to process it effectively [28].

4.4.1 Term Frequency (TF)

TF measures how frequently a term (word) appears in a document. It is calculated as the ratio of the number of times a term appears in the document to the total number of terms in that document [29].

4.4.2 Inverse Document Frequency (IDF)

IDF assesses the importance of a term across a collection of documents. Words that appear frequently across all documents are considered less important. IDF is computed as the logarithm of the total number of documents divided by the number of documents containing the term, with smoothing applied to avoid division by zero [30].

4.4.3 TF-IDF Calculation:

The TF-IDF score of a term in a document is determined by multiplying its TF (Term Frequency) and IDF (Inverse Document Frequency) values.

- We use **TfidfVectorizer** from **scikit-learn** to transform the text data into **TF-IDF vectors**. By applying this to the 'comments' column of the dataset, we generate numerical vector representations.
- The **max_features** parameter is set to 1000, limiting the feature set to the top 1000 most significant words based on their **TF-IDF scores**.

- The resulting **X-tfidf** is a sparse matrix where each row corresponds to a comment, each column represents a word, and the values indicate the **TF-IDF scores** of the words within the comments.

5 CLASSIFICATION MODEL

In this paper, we have selected the BERT and DNN models for our study, as they offer powerful capabilities in natural language processing and deep learning, respectively.

5.1 DEEP NEURAL NETWORK MODEL

DNN stands for Deep Neural Network, a type of artificial neural network characterized by multiple layers between the input and output layers [31]. These networks excel at handling complex patterns and large datasets, making them particularly powerful for advanced computational tasks [32]. DNNs are widely used in applications such as image and speech recognition, as well as natural language processing [33].

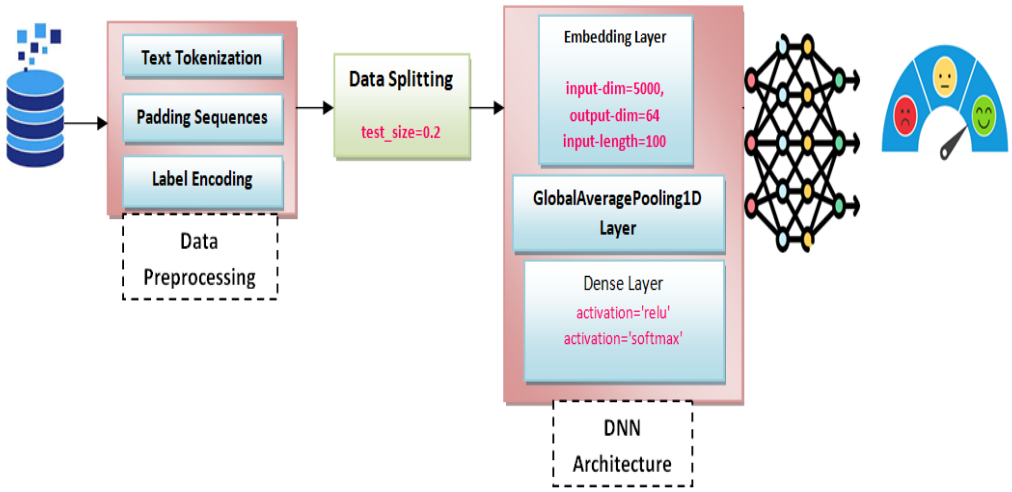


Figure 5. The DNN architecture

The Figure 5 provides a structured representation of a deep learning pipeline designed for text classification using a Deep Neural Network (DNN). It consists of two major stages: Data Preprocessing and DNN Architecture, which together form a complete workflow from raw text input to model prediction and evaluation.

1. Input Text Data:

We use pandas.read_excel() to load the data from an Excel file into a Pandas DataFrame. This allows us to handle the data in a structured format, making it easy to manipulate and process both the text and label columns.

2. Data Preprocessing:

- We use a Tokenizer to convert the text data into sequences of integers (tokens), where each unique word is assigned a numerical identifier.
- We use **pad_sequences** to ensure that all text sequences have the same length by adding padding where necessary.
- We use **LabelEncoder** to transform categorical sentiment labels (e.g. positive, neutral, and negative) into numerical values, as neural networks require numerical input for processing.

3. Data Splitting:

We use train_test_split to divide the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance. Test size = 0.2 (20% of the data is used for testing).

4. DNN Architecture:

We use an Embedding layer to convert integer-encoded words into dense vectors of fixed size. This allows the model to learn relationships between words based on their context in the training data.

- Input dimension: 5000
- Output dimension: 64
- Input length: 100

We use GlobalAveragePooling1D to reduce the output of the embedding layer by computing the average over the sequence dimension, which helps in dimensionality reduction.

We use Dense layers to perform the classification:

- The first Dense layer (with 64 units) learns complex patterns in the data using the **ReLU** activation function.
- The final Dense layer (with 3 units) outputs probabilities for each sentiment class (positive, neutral, and negative) using the **softmax** activation function.

4.2 LLM (BERT) CLASSIFIER

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model developed by Google that has transformed natural language processing (NLP) through the use of bidirectional training on a Transformer-based architecture [34]. Unlike previous models that considered text in only one direction [35], BERT examines the context of a word from both the left and right, allowing it to better understand meaning and relationships within a sentence [36]. Pre-trained on large text corpora, BERT can be fine-tuned for various NLP tasks such as sentiment analysis, question answering, and text classification [37]. Its success has inspired many variants, such as RoBERTa and DistilBERT, which optimize and extend its capabilities [38].

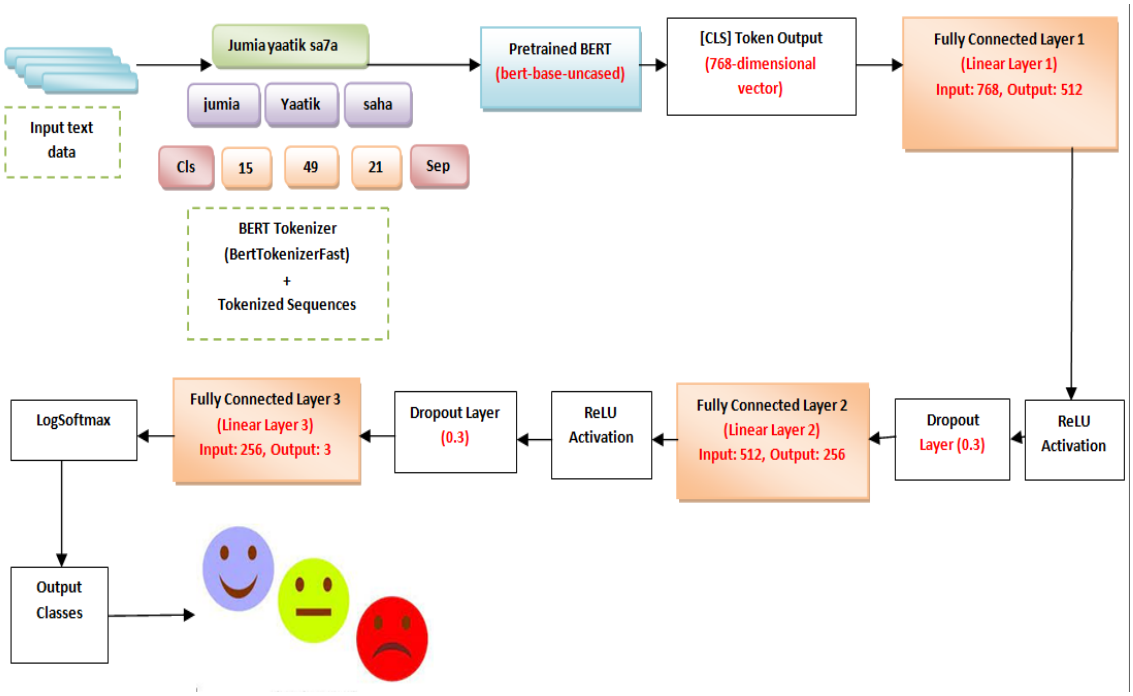


Figure 6. The General architecture of our Proposed Model using LLM (BERT model)

The different components of the architecture of our model as illustrated in Figure 6 are

Input Text Data: We utilized raw text data, which may consist of sentences, paragraphs, or entire documents, depending on the specific classification task.

BERT Tokenizer: In this step, we employed the **BertTokenizerFast** to process the text. This tokenizer divides the text into smaller units called tokens (which can be words, subwords, or characters), and each token is transformed into a numerical identifier.

Tokenized Sequences: After tokenization, we prepared the sequences for input into the BERT model. This process includes creating attention masks to indicate which tokens are relevant to the task and which are padding tokens.

Pretrained BERT: The tokenized sequences are then fed into a pretrained BERT model (bert-base-uncased). Trained on extensive unlabelled text data, this model captures rich semantic information. For each tokenized sequence, the BERT model produces a 768-dimensional vector from the [CLS] token, which serves as the representation of the entire sequence.

Fully Connected Layer 1:
The 768-dimensional vector from BERT is input into a fully connected layer with 512 output units. This layer transforms the BERT representations into a feature space better suited for the classification task.

ReLU Activation: A ReLU (Rectified Linear Unit) activation function is applied to the output of the first fully connected layer, introducing non-linearity by converting all negative values to zero.

Dropout Layer: A dropout layer is introduced after the first fully connected layer, randomly deactivating 30% of the neurons during training. This helps prevent overfitting by reducing the model's dependence on specific neural connections.

Fully Connected Layer 2: The output from the dropout layer is passed into another fully connected layer with 256 output units, performing a further linear transformation of the features.

ReLU Activation: A ReLU activation function is applied to the output of the second fully connected layer to add non-linearity to the model.

Dropout Layer: Another dropout layer is added after the second fully connected layer, again with a dropout probability of 30%.

Fully Connected Layer 3: The output of the second dropout layer is fed into a final fully connected layer with 3 output units. These units correspond to the output classes for the classification task.

LogSoftmax: To obtain class probabilities, a log softmax activation function is applied to the output of the final fully connected layer. This function converts the raw output values (logits) into logarithmic probabilities, making the results more interpretable.

Output Classes: The final output consists of logarithmic probabilities for each class, which are used to categorize the input data into one of three classes: 'positive,' 'negative,' or 'neutral.'

6. RESULTS AND DISCUTION

The evaluation phase utilizes multiple metrics to assess both the LLM (BERT) and DNN models in sentiment classification. Accuracy provides an overall measure of correct predictions across the dataset. The F1-score balances precision and recall, offering a comprehensive view of performance.

Additional metrics like precision, recall, and the confusion matrix offer deeper insights into model behavior. These evaluations highlight both models' strengths and areas needing improvement. Together, this multi-metric approach ensures a robust and effective assessment of model performance.

6.1 DNN RESULTS

The results shown in Table 2 were obtained using the DNN model on our dataset.

Table. 2 DNN Results

	Epoch=20	**Epoch=30**	**Epoch=40**	**Epoch=50**
Accuracy	0,64	0,77	0,86	0,75
Precision	0,82	0,85	0,80	0,82
Recall	0,78	0,82	0,83	0,7
F1-score	0,69	0,75	0,78	0,84

The results indicate that the DNN model's performance evolves significantly over the epochs, with the best overall performance observed at different stages for various metrics. Here is a detailed analysis:

Epoch 40 – Peak Accuracy and Recall: At epoch 40, the model achieves its highest accuracy of 86% and the best recall at 83%. These figures suggest that the model is highly effective at correctly classifying instances and capturing most of the relevant cases. High recall is particularly important in scenarios where missing a relevant case

could be critical. Although the F1-score at this stage is 78%, the elevated accuracy and recall highlight that the model is performing well in recognizing true positives.

Epoch 50 – Best F1-Score: By epoch 50, the model's F1-score reaches its maximum value of 84%, which indicates a strong balance between precision and recall. This improvement suggests that, despite a slight drop in accuracy (75%) and recall (70%), the overall balance between false positives and false negatives has been optimized. The F1-score is especially valuable in sentiment classification tasks where both precision (avoiding false alarms) and recall (capturing true instances) are crucial.

Consistency in Precision: Across all epochs, precision remains relatively stable (ranging between 0.80 and 0.85). This consistency implies that the model reliably classifies positive instances correctly without a significant number of false positives, regardless of the training epoch.

Trade-Offs and Model Behavior: The difference between the highest accuracy and recall at epoch 40 versus the best F1-score at epoch 50 suggests a trade-off inherent in the training process. While epoch 40 shows that the model is highly capable of identifying most relevant instances, epoch 50 indicates that fine-tuning further improves the balance between precision and recall. Such trade-offs are common in machine learning, where extending training can lead to overfitting on some metrics while enhancing others.

In summary, while epoch 40 stands out for achieving peak accuracy and recall, the best overall balance between precision and recall, as measured by the F1-score, is observed at epoch 50. This analysis helps in identifying the stage at which the model's performance is most robust for the specific requirements of the sentiment classification task. These advancements are vividly depicted in the histogram visualization presented in Figure 7, showcasing the model's steady progress.

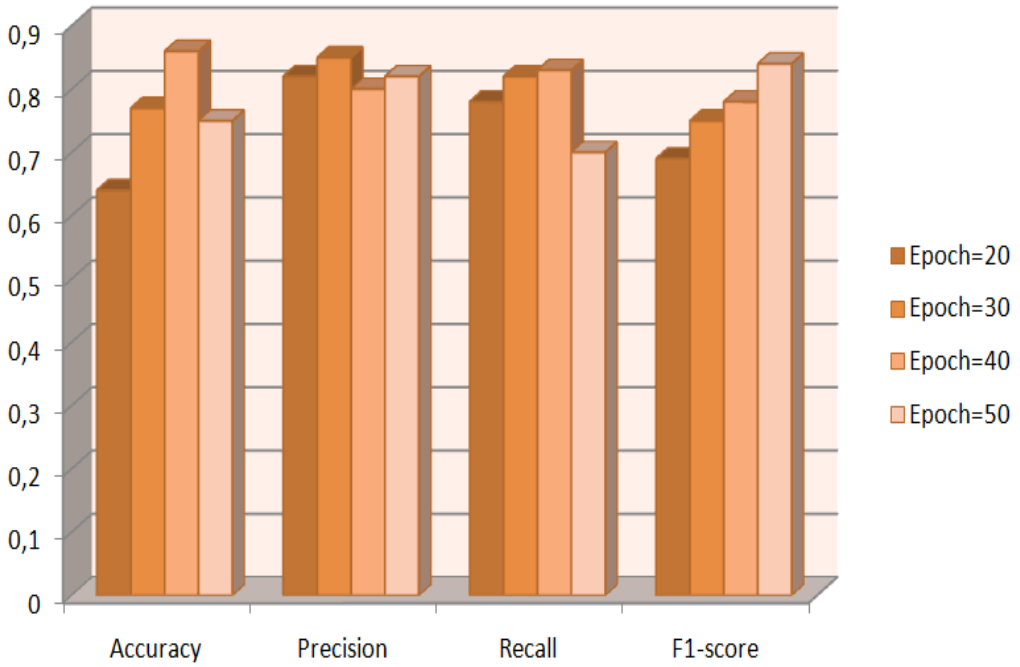


Figure 7. Histogram represents the development of results using DNN.

6.2 BERT RESULTS

Table 1 presents the performance metrics of a BERT (Bidirectional Encoder Representations from Transformers) model across four different runs, each trained for 30 epochs. The metrics include Accuracy, Precision, Recall, and F1-score, which are commonly used to evaluate the performance of classification models.

Table. 3 BERT Results

	Epoch=20	**Epoch=30**	**Epoch=40**	**Epoch=50**
Accuracy	0,9544	0,9519	0,9494	0,9519
Precision	0,931	0,93	0,9297	0,928
Recall	0,9557	0,9557	0,9531	0,9565
F1-score	0,947	0,9387	0,9369	0,938

The results presented in Table 3 were obtained using the LLM (BERT model) on our dataset, yielding exceptional performance that surpassed that of the DNN model. Remarkably, when training for different numbers of epochs (20, 30, 40, 50), the BERT model exhibited its best performance at 20 epochs, after which a gradual decline in performance was observed. At 20 epochs, the model achieved peak values for key performance metrics, with both accuracy and precision reaching their highest levels. In addition, recall and the F1-score—representing the harmonic mean of precision and recall—showed significant improvements, as illustrated in Figure 8. This indicates that

the model was not only precise but also highly effective at identifying relevant instances, further underscoring the superior capability of the BERT model compared to the DNN model.

Specifically, as detailed in Table 1 (BERT Results), at 30 epochs, the BERT model consistently achieved an accuracy of approximately 95%, precision around 93%, recall close to 95%, and an F1-score of about 94%. These impressive results confirm that the model maintains robust performance across multiple evaluation criteria.

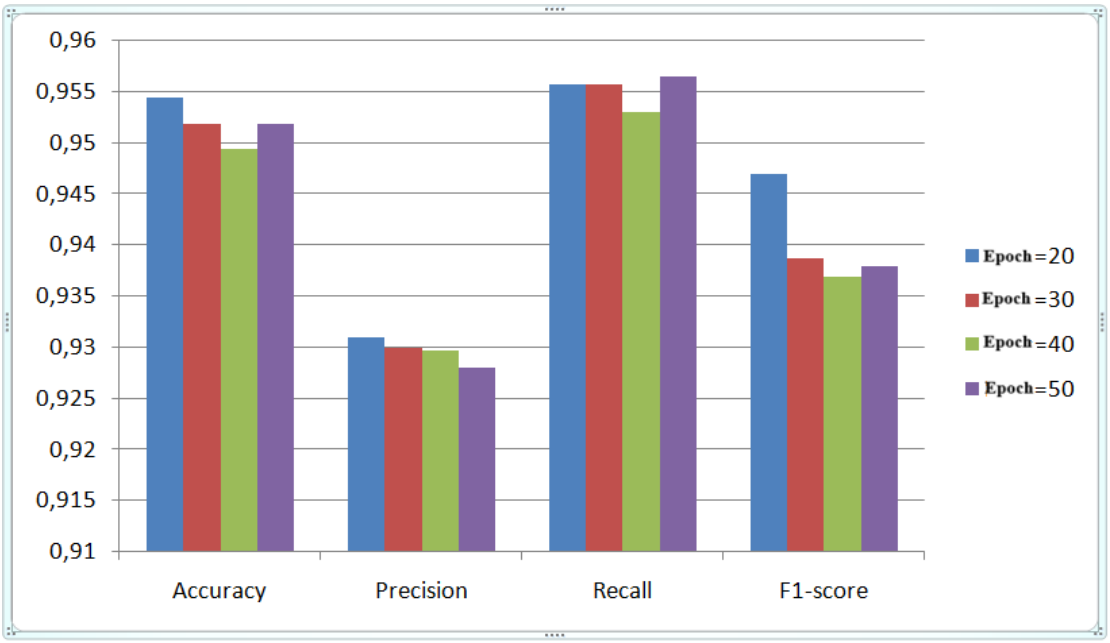


Figure 8. Histogram represents the development of results using LLM (BERT model).

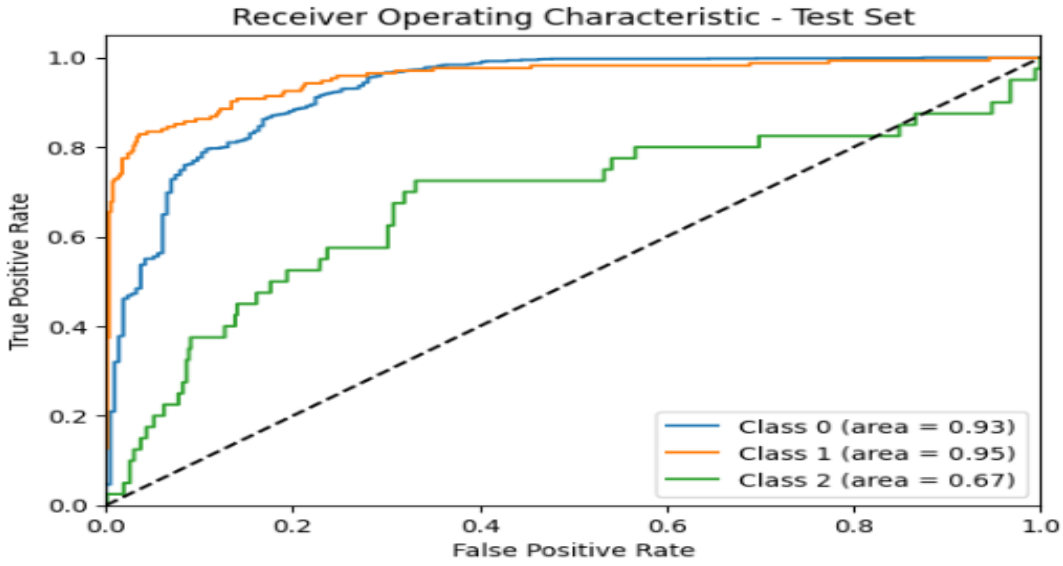


Figure 9. ROC curve Results for BERT

Furthermore, the ROC curve shown in Figure 9 demonstrates excellent discriminative ability, highlighting the model's robustness in distinguishing between sentiment classes. Figure 10 provides a clear comparison of accuracy between the DNN and BERT models, emphasizing the BERT model's superior performance, especially in processing data in the Algerian dialect.

These results not only affirm the efficacy of using pre-trained transformer-based models like BERT for sentiment classification but also highlight the importance of fine-tuning training epochs to achieve optimal performance. The ability of BERT to capture rich semantic information and maintain high performance across multiple metrics positions it as a highly promising approach for enhancing analysis systems. Consequently, this research lays a robust foundation for further exploration into transformer-based architectures, offering potential improvements for various natural language processing applications and optimization techniques in diverse fields.

6.3 Generalizability and Applicability

While our experiments have been conducted using data from the Jumia platform, the proposed methodology is designed with broader applicability in mind. The underlying framework comprising GAN-based data augmentation and the application of advanced deep learning models (BERT and DNN) is not inherently limited to Jumia. Instead, it offers a flexible approach that can be adapted to a variety of contexts and data sources with similar linguistic challenges.

In practice, the transferability of our approach to other e-commerce platforms or social media sources hinges on domain adaptation techniques. For instance, by fine-tuning the BERT model and adjusting data pre-processing strategies, our system can be calibrated to accommodate differences in sentiment expression, text length, and domain-specific vocabulary that may occur on platforms such as Facebook, Twitter, or other online retail sites.

Moreover, our GAN-based data augmentation method enhances the model's robustness by simulating diverse linguistic scenarios. This not only enriches the training corpus for the Algerian dialect but also provides a blueprint for extending the approach to other low-resource dialects or languages. Although empirical testing on these additional

datasets remains a future direction, preliminary evaluations suggest that our methodology can maintain high performance levels when adapted to new data contexts.

Overall, these considerations underscore the potential for broader applicability of our findings and encourage further research to validate and refine the model's performance across different domains.

6.4 Limitations

While our study demonstrates promising results for sentiment analysis in the Algerian dialect using data from the Jumia platform, several limitations should be highlighted:

- **Dialect and Linguistic Variability**: This study focuses solely on the Algerian dialect. Although our approach addresses some challenges associated with low-resource dialects, the effectiveness of the models might vary when applied to other dialects or languages with different linguistic characteristics.
- **Data Augmentation Challenges**: While the GAN-based data augmentation technique enriched the dataset, the synthetic data generated may not fully capture the subtleties and natural variability of real-world language. This could introduce noise or bias into the training process, affecting model performance.
- **Annotation Subjectivity**: The manual annotation of user reviews, despite careful execution, is inherently subjective. Variations in annotator interpretation can lead to inconsistencies in labelling, which may influence the accuracy of the sentiment classification.
- **Model Scalability and Computational Resources**: The deep learning models (BERT and DNN) were optimized for our current dataset. Their scalability, as well as the computational resources required for training and fine-tuning on larger and more diverse datasets, remains to be fully explored.
- **Hyperparameter Tuning and Model Complexit**y: Although extensive experiments were performed, further refinement through comprehensive hyperparameter tuning and exploration of alternative model architectures could potentially enhance performance.

Recognizing these limitations provides a balanced perspective on our research and lays the groundwork for future investigations aimed at improving and generalizing sentiment analysis methodologies for diverse linguistic contexts.

7. CONCLUSION

In conclusion, this research developed a sentiment analysis model tailored to the Algerian dialect, utilizing data collected from the Jumia website. We began by constructing a specialized dictionary to capture the nuances of the Algerian dialect, a critical step in addressing the region-specific language variations. This enabled us to create a sentiment analysis framework capable of accurately interpreting the context and sentiments expressed in the local dialect.

We employed two advanced deep learning techniques: BERT (Bidirectional Encoder Representations from Transformers) and DNN (Deep Neural Network). BERT, a state-of-the-art language model, proved particularly effective due to its ability to understand the context of words bidirectionally, making it highly suitable for sentiment analysis tasks. On the other hand, DNN provided a robust framework for identifying complex patterns in the data, enhancing the accuracy of sentiment classification.

Throughout the project, we adopted a systematic approach, encompassing data preprocessing, model training, and performance evaluation. These steps were essential to ensure the quality and reliability of our results. During model training, we experimented with various hyperparameters and fine-tuned the models to optimize their performance.

The final results demonstrated the superior effectiveness of the BERT-based language model for sentiment analysis in the Algerian dialect. BERT's contextual understanding, combined with DNN's ability to recognize intricate data patterns, resulted in high accuracy and performance in sentiment classification. This research underscores the potential of deep learning techniques and advanced language models like BERT in addressing the unique challenges of dialectal language processing.

8. REFERENCES

[1] S. Mdhaffar, F. Bougares, Y. Esteve, and L. Hadrich-Belguith, "Sentiment analysis of Tunisian dialects: Linguistic resources and experiments," in *Third Arabic Natural Language Processing Workshop*, 2017: ACL, pp. 55–61.

[2] A. Chader, D. Lanasri, L. Hamdad, M. C. E. Belkheir, and W. Hennoune, "Sentiment analysis for Arabizi: Application to Algerian dialect," in *11th*

International Conference on Knowledge Discovery and Information Retrieval, 2019: SCITEPRESS, pp. 475–482.

[3] O. Boucherit and K. Abainia, "Offensive language detection in under-resourced Algerian dialectal Arabic language," in *International Conference on Big Data, Machine Learning, and Applications*, 2021: Springer, pp. 639–647.

[4] I. Guellil et al., "A semi-supervised approach for sentiment analysis of Arab(ic+izi) messages: Application to the Algerian dialect," *SN Computer Science*, vol. 2, pp. 1–18, 2021.

[5] B. Klouche and S. Benslimane, "Arabizi chat alphabet transliteration to Algerian dialect," in *International Conference in Artificial Intelligence in Renewable Energetic Systems*, 2020: Springer, pp. 790–797.

[6] K. Z. Bousmaha, K. Hamadouche, I. Gourara, and L. B. Hadrich, "DZ-OPINION: Algerian dialect opinion analysis model with deep learning techniques," *Revue d'Intelligence Artificielle*, vol. 36, no. 6, p. 897, 2022.

[7] A. Soumeur, M. Mokdadi, A. Guessoum, and A. Daoud, "Sentiment analysis of users on social networks: Overcoming the challenge of the loose usages of the Algerian dialect," *Procedia Computer Science*, vol. 142, pp. 26–37, 2018.

[8] H. Houari and M. Guerti, "Study the influence of gender and age in recognition of emotions from Algerian dialect speech," *Traitement du Signal*, vol. 37, no. 3, pp. 413-423, 2020.

[9] I. Guellil, M. Mendoza, and F. Azouaou, "Arabic dialect sentiment analysis with ZERO effort: Case study: Algerian dialect," *Inteligencia Artificial*, vol. 23, no. 65, pp. 124–135, 2020.

[10] J. Younes, H. Achour, E. Souissi, and A. Ferchichi, "A deep learning approach for the Romanized Tunisian dialect identification," *International Arab Journal of Information Technology*, vol. 17, no. 6, pp. 935–946, 2020.

[11] M. A. Menacer and K. Smaïli, "Investigating data sharing in speech recognition for an under-resourced language: The case of Algerian dialect," in *7th International Conference on Natural Language Processing (NATP)*, 2021, pp. 77–89.

[12] K. Lounnas, M. Abbas, and M. Lichouri, "Towards phone number recognition for code-switched Algerian dialect," in *4th International Conference on Natural Language and Speech Processing (ICNLSP)*, 2021, pp. 290–294.

[13] S. Bougrine, H. Cherroun, and D. Ziadi, "Hierarchical classification for spoken Arabic dialect identification using prosody: Case of Algerian dialects," *arXiv preprint* arXiv:1703.10065, 2017.

[14] M. A. Menacer et al., "Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect," *Procedia Computer Science*, vol. 117, pp. 81–88, 2017.

[15] M. Lichouri, M. Abbas, A. A. Freihat, and D. E. H. Megtouf, "Word-level vs sentence-level language identification: Application to Algerian and Arabic dialects," *Procedia Computer Science*, vol. 142, pp. 246–253, 2018.

[16] M. A. Cheragui, A. H. Dahou, and A. Abdedaiem, "Exploring BERT models for part-of-speech tagging in the Algerian dialect: A comprehensive study," in *6th International Conference on Natural Language and Speech Processing (ICNLSP)*, 2023: IEEE, pp. 140–150.

[17] C. Garvey and C. Maskal, "Sentiment analysis of the news media on artificial intelligence does not support claims of negative bias against artificial intelligence," *OMICS: A Journal of Integrative Biology*, vol. 24, no. 5, pp. 286–299, 2020.

[18] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 33, 2021.

[19] Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the application of deep learning-based BERT model in sentiment analysis," *arXiv preprint* arXiv:2403.08217, 2024.

[20] Y. Cao, Z. Sun, L. Li, and W. Mo, "A study of sentiment analysis algorithms for agricultural product reviews based on improved BERT model," *Symmetry*, vol. 14, no. 8, p. 1604, 2022.

[21] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment analysis of Persian movie reviews using deep learning," *Entropy*, vol. 23, no. 5, p. 596, 2021.

[22] V. Umarani, A. Julian, and J. Deepa, "Sentiment analysis using various machine learning and deep learning techniques," *Journal of the Nigerian Society of Physical Sciences*, pp. 385–394, 2021.

[23] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big data: Deep learning for financial sentiment analysis," *Journal of Big Data*, vol. 5, no. 1, pp. 1–25, 2018.

[24] J. Wang, B. Xu, and Y. Zu, "Deep learning for aspect-based sentiment analysis," in *International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, 2021: IEEE, pp. 267–271.

[25] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, p. 652801, 2021.

[26] H. S. Muti et al., "The Aachen protocol for deep learning histopathology: A hands-on guide for data preprocessing," *Zenodo*, 2020.

[27] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast wordpiece tokenization," *arXiv preprint* arXiv:2012.15524, 2021.

[28] A. P. Wibawa, H. K. Fithri, I. A. E. Zaeni, and A. Nafalski, "Generating Javanese stopwords list using K-means clustering algorithm," *Knowledge Engineering and Data Science*, vol. 3, no. 2, pp. 106–111, 2020.

[29] S. Amin et al., "Recurrent neural networks with TF-IDF embedding technique for detection and classification in tweets of dengue disease," *IEEE Access*, vol. 8, pp. 131522–131533, 2020.

[30] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "LSTM, VADER and TF-IDF based hybrid sentiment analysis model," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.

[31] M. Ahmed, Q. Chen, Y. Wang, Y. Nafa, Z. Li, and T. Duan, "DNN-driven gradual machine learning for aspect-term sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, 2021: ACL, pp. 488–497.

[32] R. Wadawadagi and V. Pagi, "Sentiment analysis with deep neural networks: Comparative study and performance assessment," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 6155–6195, 2020.

[33] J. Khan, N. Ahmad, S. Khalid, F. Ali and Y. Lee, "Sentiment and Context-Aware Hybrid DNN With Attention for Text Sentiment Classification," *IEEE Access*, vol. 11, pp. 28162-28179, 2023.

[34] M. Masala, S. Ruseti, and M. Dascalu, "ROBERT – A Romanian BERT model," in *28th International Conference on Computational Linguistics (COLING)*, 2020: ICCL, pp. 6626–6637.

[35] J. Cañete et al., "Spanish pre-trained BERT model and evaluation data," *arXiv preprint* arXiv:2308.02976, 2023.

[36] L. Akhtyamova, "Named entity recognition in Spanish biomedical literature: Short review and BERT model," in *26th Conference of Open Innovations Association (FRUCT)*, 2020: IEEE, pp. 1–7.

[37] H. Lee, S. Lee, I. Lee, and H. Nam, "AMP-BERT: Prediction of antimicrobial peptide function based on a BERT model," *Protein Science*, vol. 32, no. 1, p. e4529, 2023.

[38] T. ValizadehAslani et al., "PharmBERT: A domain-specific BERT model for drug labels," *Briefings in Bioinformatics*, vol. 24, no. 4, p. bbad226, 2023.