الجمهورية الجزائرية الديمقراطية الشعبية
The People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة محمد بوضياف بالمسيلة
University Mohamed Boudiaf of M'sila

كلية الرياضيات والإعلام الآلي
Faculty of Mathematics and Informatics

قسم الإعلام الآلي
Department of Computer Science

Domain: Mathematics and Computer Science

Thesis Presented to Fulfill the Partial Requirement
for **Master's Degree** in Computer Science

Specialty: Networks and information and communication technologies

Prepared By: Reguig Berra Haithem

Supervised By:

Bentercia Rahima

Entitled

Sentiment Analysis of Tweets Related to the Gaza War

Jury Members

Dr Abdallah tharafi	President
Dr.Rahima bentrcia	Supervisor
Dr.Marouane kihal	Examiner

Academic Year 2024/2025

Dedication

I dedicate this humble work to those who shaped my journey and lifted me to this level of knowledge. Allah the Exalted said:

﴿ وَاخْفِضْ لَهُمَا جَنَاحَ الذُّلِّ مِنَ الرَّحْمَةِ وَقُل رَّبِّ ارْحَمْهُمَا كَمَا رَبَّيَانِي صَغِيرًا﴾

To My Beloved Mother

Your love, strength, and unwavering support have been my foundation. You stood beside me through every challenge, prayed for me through every step, and believed in me even when I doubted myself. This work is a tribute to your sacrifices, your patience, and the light you bring into my life every day. I am endlessly grateful, and I carry your love with pride and honor.

To the Memory of My Dear Father

Though you are no longer with us, your presence remains alive in my heart. Your wisdom, values, and silent sacrifices continue to guide me. This achievement is as much yours as it is mine, and I hope to make you proud in every step I take.

To My Siblings

Thank you for the joy, laughter, and support you bring. Your encouragement has meant the world to me, and your presence continues to enrich my life.

Reguig Berra Haithem

Acknowledgement

First and foremost, all praise and gratitude are due to **Allah, the Most Gracious and the Most Merciful**, who blessed us with the strength, patience, and determination to complete this work. Without His divine guidance and mercy, none of this would have been possible.

Our deepest appreciation goes to **Dr. Rahima Bentrcia**, whose unwavering support, insightful supervision, and generous investment of time and knowledge played a pivotal role in shaping the success of this thesis. Her encouragement and guidance have been a source of continuous motivation, and her constructive critiques have sharpened our vision and enriched the quality of our research. We are truly privileged and honored to have had her as our supervisor, and we will remain forever grateful for her mentorship and kindness throughout this journey.

الملخص

يهدف هذا البحث إلى تحليل المشاعر في التغريدات المكتوبة باللغة العربية المتعلقة بصراع غزة. تم بناء مجموعة البيانات من خلال جمع تغريدات من منصات التواصل الاجتماعي ودراسات سابقة، مما يضمن تنوعًا في اللهجات ووجهات النظر. خضعت التغريدات لعملية معالجة مسبقة باستخدام كود برمجي مخصص بلغةPython ، حيث شملت إزالة الرموز التعبيرية، وتوحيد الصيغ الإملائية، وحذف الكلمات الشائعة غير المؤثرة. بعد ذلك، تم تصنيف المشاعر يدويًا على أنها إيجابية، سلبية أو محايدة. تم تدريب نموذج AraBERT معدل لأداء مهمة التصنيف بدقة. يسمح النظام النهائي بإدخال تغريدات جديدة باللغة العربية وتحليل مشاعرها بدقة عالية.

الكلمات المفتاحية :
معالجة اللغة الطبيعية، التصنيف AraBERT تحليل المشاعر، اللغة العربية، التغريدات، صراع غزة، التعلم العميق.

abstract

This research aims to analyze the sentiment of Arabic tweets related to the Gaza conflict. The dataset was built by collecting tweets from social media platforms and previous studies, ensuring a diversity of dialects and perspectives. The tweets underwent preprocessing using a custom Python code, which included removing emojis, standardizing spellings, and removing ineffective common words. Sentiments were then manually classified as positive, negative, or neutral. A modified AraBERT model was trained to accurately perform the classification task. The final system allows for the input of new Arabic tweets and their sentiment analysis with high accuracy.

Key Words:
Sentiment Analysis, Arabic Language, Tweets, Gaza Conflict, Deep Learning, AraBERT, NLP, Classification.

Résumé

Cette recherche vise à analyser le sentiment dans les tweets arabes liés au conflit de Gaza. L'ensemble de données a été construit en collectant des tweets provenant de plateformes de médias sociaux et d'études antérieures, garantissant une diversité de tons et de perspectives. Les tweets ont été prétraités à l'aide d'un code Python personnalisé, qui comprenait la suppression des emojis, la normalisation de l'orthographe et la suppression des mots courants inefficaces. Ensuite, les émotions ont été classées manuellement comme positives, négatives ou neutres. Un modèle AraBERT modifié est formé pour effectuer avec précision la tâche de classification. Le

système final permet la saisie de nouveaux tweets en arabe et leur analyse des sentiments avec une grande précision

Mots-clés :

Analyse de sentiments, langue arabe, tweets, conflit de Gaza, apprentissage profond, AraBERT, traitement automatique du langage, classification.

Table of Contents

Contents

List of Figures

List of Tables

General Introduction

This task is aimed at developing an Arabic social media sentiment analysis system that would be focused on sensitive and emotive topics such as the Gaza war. Sentiment analysis for Arabic is difficult due to the fact that the language has a complex morphology, plentiful vocabulary, and also because there exists a wide use of different dialects across areas. Moreover, emotional expressions in Arabic, especially in politics and humanitarian cases, are intricate and culturally rooted, so that an overall-purpose sentiment analysis system would not necessarily be able to gauge them properly. This would previously guide the available models to mislabel or oversimplify the actual sentiment, particularly in politically or socially sensitive circumstances.

In addressing this void, our system draws upon Arabic Natural Language Processing (NLP) techniques and employs a fine-tuned AraBERT model that draws upon contextual word embeddings in order to detect subtle sentiment triggers more effectively. Analysis is focused on categorizing tweets into three broad categories: positive, negative, and neutral.

The project follows a systematic process with the following significant steps: data extraction (Twitter API, literature, and social media websites), data preprocessing (cleaning, normalization, and dialect processing), manual sentiment tagging, model training, and performance metrics.

The main problem that this research is attempting to address is the lack of proper and contextually aware sentiment analysis resources for the Arabic language, especially when applied to actual scenarios such as the Gaza crisis. Existing models do not generalize well across dialects or recognize the socio-political implications contained in Arabic text, making for skewed or inaccurate readings of public sentiment.

The objectives of this research are as follows:

To build a large-scale and representative Arabic corpus of tweets regarding the Gaza crisis.

To design an effective data preprocessing pipeline specific to the nature of the Arabic language

To fine-tune and apply a transformer model (AraBERT) to achieve sentiment classification precision.

To evaluate the system performance using suitable metrics such as accuracy, F1-score, precision, and recall

To understand the emotional responses of the masses towards the ongoing Gaza crisis using social media.

This thesis is organized as follows:

Chapter 1 is a literature review, outlining the main methods and current research in Arabic sentiment analysis and dialect and emotional usage problems.

Chapter 2 presents the theoretical background, discussing important NLP terms, transformer models, and sentiment classification.

Chapter 3 presents the methodology, including data sources, preprocessing steps, and training model procedures.

Chapter 4 outlines experimental findings, evaluation metrics, and discussion of results.

Finally, the conclusion outlines the main contributions, shortcomings of the project, and also possible directions for future research.

Through addressing the unique linguistic and cultural features of Arabic text, this project attempts to help make sentiment analysis tools more inclusive and accurate, particularly when handling humanitarian crises.

CHAPTER 1 : Arabic Sentiment Analysis in Social Media Contexts

1. Introduction

In recent years, the fast growth of social media platforms has made them a basic means for people to express their ideas, opinions, and emotions, especially on sensitive and politically charged topics. The Gaza conflict, in particular, sparked widespread discussions across these platforms, with individuals expressing strong emotions and diverse opinions in response to the unfolding events.

However, explaining these sentiments can be challenging, especially when dealing with a language as diverse and linguistically complex as Arabic. The richness of the Arabic language, coupled with its various dialects and the subtle nuances in which emotions is conveyed, presents unique obstacles for sentiment analysis. Traditional sentiment analysis models often fail to capture the exact expressions found in language, leading to inaccurate or incomplete interpretations of public sentiment.

Recent advances in natural language processing have led to the development of more sophisticated technologies, such as models. One of the most promising models for Arabic sentiment analysis is AraBERT, a variant of the widely-used BERT architecture that has been specifically tuned for the Arabic language.

This project aims to leverage these developments to analyze the public sentiment surrounding the Gaza conflict on social media platforms. By utilizing the AraBERT model and other advanced NLP techniques, we aim to develop a robust sentiment analysis system capable of classifying Arabic text into categories such as positive, negative, and neutral, based on feelings conveyed in social media content. This sentiment classification can provide valuable insights into public opinion, which can help guide humanitarian efforts, inform political decision-making, and contribute to a deeper understanding of the societal impact of the conflict.

The project follows several major phases: collection of data, preprocessing of data, training and tuning of models, and evaluation. Our goal here is to improve the accuracy of sentiment analysis for Arabic social media, especially in the context of emotively charged events such as the Gaza conflict. Ultimately, it is envisioned that this project would provide an appreciable advancement toward sonar knowledge in the area such that the knowledge gleaned could provide insights that may assist in acting on such crises.

2. Problem Statement

Accurately analyzing and classifying Arabic language content on social media poses

significant challenges due to the diversity of Arabic dialects and nuanced expressions. This project will attempt to address these issues in the following manner:

Addressing Arabic Linguistic Diversity: Arabic is a complex language with diverse regional dialects, unique sentence structures, and extensive vocabulary, making sentiment analysis difficult for traditional models.

Understanding Subtle Sentiment Expressions: Social media users express emotions in indirect or subtle ways, especially when sensitive topics are involved.

Emotionally Charged Content Interpretation: The Gaza conflict is a very sensitive and controversial topic. These challenges are heightened in the interpretation of public sentiment, which is usually very emotional such crises and requires more sophisticated tools.

This project employs the latest in natural language processing, especially the AraBERT model, to conquer these challenges. In this work, we build a robust sentiment analysis system that accurately classifies Arabic text into sentiment categories: positive, negative, or neutral. This is addressed with the linguistic diversity and rich subtlety in the expression of emotions peculiar to Arabic.

3. Challenges Related to Arabic Sentiment Analysis

Arabic sentiment analysis poses a unique set of challenges that stem from the linguistic and cultural complexities of the language, particularly when analyzing content from diverse social media platforms. These challenges include:

3.1 Diverse Dialects and Regional Variations:

Arabic is one of the most diverse languages, with multiple dialects such as Egyptian Arabic, Levantine Arabic, Gulf Arabic and Maghrebi Arabic, along with Modern Standard Arabic (MSA). Such dialects are very divergent at the level of vocabulary, syntax and grammar which make difficult for sentiment analysis models to give consistent and accurate results.

Dialect/Region	Expression in Local Dialect	Expression in MSA	Meaning in English
Egyptian	عامل إيه؟	كيف حالك؟	How are you?
Algeria	واش راك ؟	كيف حالك؟	How are you?
Saudi Arabian	وش اخبارك ؟	كيف حالك؟	How are you?
Iraqi	شلونك؟	كيف حالك؟	How are you?
Lebanese	كيفك؟	كيف حالك؟	How are you?

Table 1: Examples of Diverse Arabic Dialects in Sentiment Analysis.

3.2 Code-Switching

Many Arabic speakers mix Arabic with other languages, such as English or French, within a single sentence or post. This phenomenon, known as code-switching, adds an extra layer of complexity for sentiment analysis models, which must handle multiple languages simultaneously.

Type	Expression with Code-Switching	Fully in Arabic Fully in Arabic	Meaning in English
Arabic-English	خلصت الميتنج قبل شوية	انتهيت من الاجتماع قبل قليل	I just finished the meeting
Arabic-French	الدرس كان تري بيان	الدرس كان جيدًا جدً	The lesson was very good

Table 2: Impact of Code-Switching on Sentiment Analysis.

3.3 Ambiguity and Context Dependency

Arabic expressions often carry multiple meanings depending on the context in which they are used. Sarcasm, idiomatic phrases, and culturally specific references are common, further complicating the interpretation of sentiment. As illustrated in Figure 1.1, an example of an ambiguous tweet is shown, where a metaphor about hope in difficult times is used. This tweet is ambiguous in determining whether it conveys hope effectively, highlighting the challenge of interpreting sentiment in such nuanced contexts [1].

Figure 1 : An example of Ambiguity tweet.

3.4 Informality of Social Media Language:

Social media platforms are dominated by informal language, which includes slang, abbreviations, emojis, and unconventional spellings. These features are particularly prevalent in Arabic tweets and posts, making it difficult to tokenize and analyze the text using traditional NLP techniques.

3.5 Emotionally Charged and Sensitive Topics:

Topics like the Gaza conflict tend to evoke intense emotions, which are often expressed in subtle or highly nuanced ways. Capturing these emotions accurately requires advanced models capable of understanding context and emotional intensity.

3.6 Linguistic Features Unique to Arabic:

Features such as rich morphology, right-to-left script, and the absence of standard orthography in dialectal Arabic complicate the preprocessing and tokenization stages of sentiment analysis.

4. Novel Contributions of this Thesis

This thesis presents several innovative contributions to the field of Arabic sentiment analysis and natural language processing:

4.1 Publicly Available Dataset and Model

This project offers a carefully curated dataset of Arabic tweets concerning the Gaza conflict, annotated for sentiment, along with a fine-tuned version of the AraBERT model. These resources are provided to support future research and applications in Arabic NLP.

4.2 Integration of Diverse Arabic Dialects in Sentiment Analysis

This study tackles the issue of dialectal differences by incorporating tweets and posts from different Arabic-speaking areas. The developed model shows enhanced effectiveness in interpreting and categorizing sentiment in both Modern Standard Arabic (MSA) and various dialects.

4.3 Insights into Public Sentiment on the Gaza Conflict

This thesis analyzes the public's emotional reactions during significant events of the Gaza conflict by applying the developed model. The findings provide important insights for humanitarian organizations, policymakers, and researchers examining the societal effects of political crises.

5. conclusion

This chapter highlighted the complexities of Arabic sentiment analysis, focusing on the unique linguistic challenges posed by the Arabic language and its diverse dialects. It also shed light on the subtleties of sentiment expression, especially in sensitive contexts such as the Gaza conflict, where public opinions are deeply emotional and multifaceted.

The organization of this thesis ensures a logical flow, starting with starting from the introduction of the problem and challenges, followed by data set introduction, literature review, and detailed explanation of

the proposed system. The results and analysis chapter provides insights into the system's performance, and the conclusion chapter summarizes the research and discusses future directions

Chapter 2: Literature Review

1. Introduction:

Sentiment analysis is now an established field of research on human emotions, attitudes, and opinions expressed through text—most noticeably on social media. As data available on the web grows, sentiment analysis of user-generated content has gained importance in a number of domains, including marketing, politics, journalism, and humanitarian studies.

While sentiment analysis in English has been extensively studied and aided by numerous datasets and tools, the state of Arabic sentiment analysis is still in its infancy. Arabic language has additional difficulties with its complex morphology, rich syntax, and the presence of a number of regional dialects that contrast significantly from Modern Standard Arabic (MSA). These linguistic features require additional complex approaches to adequately extract meaning, tone, and context.

Over the past decade, researchers have explored various methods to tackle sentiment analysis in Arabic, ranging from traditional machine learning techniques to deep learning and hybrid approaches. Early studies focused on manually engineered features and statistical models, but recent advancements in natural language processing (NLP)—especially the introduction of transformer-based models like BERT and its Arabic variants such as AraBERT—have significantly improved the performance of sentiment classification systems.

This chapter provides a comprehensive review of pertinent studies in Arabic sentiment analysis. It begins with an overview of traditional machine learning methods, is then followed by an explanation of the innovation brought about by deep learning architectures, and finally touches on hybrid ones that aim to get the best out of the two. The purpose of this review is to highlight the advancement in techniques utilized, create gaps in previous studies, and elucidate the reason pre-trained transformer models—AraBERT in particular—are utilized in this study for analyzing Arabic tweets regarding Gaza conflict.

2. Lexicon-Based Approach

The lexicon-based approach to sentiment analysis relies on predefined dictionaries (lexicons) containing sentiment-bearing words annotated with their corresponding polarities (positive,

negative, or neutral). When analyzing a text, the sentiment is computed by aggregating the scores of individual words. This method does not require labeled training data, making it particularly useful for low-resource languages and dialects. In Arabic sentiment analysis, several lexicons have been developed, either manually or automatically, and applied across different dialects [2].

Various studies have used this approach with promising results. For instance, accuracy levels above 85% were reported on Egyptian and Gulf Arabic texts. The method has also been tested on social media platforms like Facebook and Twitter, proving effective despite dialectal and orthographic variations.

Ref	Dataset Size & Source	Dialect	Labels	Performance
[3]	3,484 – Facebook Comments	Egyptian	Pos, Neg	98.20% Acc (Pos), 93.20% Acc (Neg)
[4]	4,700 – Twitter Posts	Gulf	Pos, Neg	85.40% Accuracy
[5]	1,500 – Twitter Posts	Gulf	Pos, Neg, Neu	F1: 77.4% (Pos), 59.1% (Neg), 51.1% (Neu)
[5]	7,698 – Facebook Comments	Algerian	Pos, Neg, Neu	79.13% Accuracy
[6]	2,000 – Hotel Reviews	MSA	Pos, Neg	91.0% Accuracy

Table 3:Accuracy and F1 Scores of Sentiment Analysis Across Arabic Dialects

3. Deep Learning Approach

This approach relies on the utilization of labeled datasets for training classification algorithms so that they automatically learn how to identify the sentiment of a text. Unlike lexicon-based approaches, machine learning (ML) models learn complex patterns from features such as term frequency, n-grams, or word embeddings. Common algorithms used are Support Vector Machines (SVM), Naive Bayes, Decision Trees, Random Forests, and more recently deep learning models such as CNN, LSTM, and BERT variants.

They require much labeled data to train and perform and tend to do better than lexicon-based methods, particularly on dialectal Arabic, sarcasm, and uncertain sentiment [7].

Reference	Dataset Size & Source	Dialect	Labels	ML Model	Performance
[8]	21,000 tweets (Twitter)	MSA	Pos, Neg	SVM, Naive Bayes	Acc: 90.5%
[9]	17,573 tweets (AraSenTi-Tweet)	Saudi (Gulf)	Pos, Neg, Neu	Decision Trees	F1-score: 81.5%
[10]	250K tweets (ASTD)	Egyptian	Pos, Neg, Neu	SVM, RF, NB	Acc: 77.5%
[11]	10,000 tweets (Twitter)	Jordanian	Pos, Neg	Logistic Regression	Acc: 85.0%

Table 4: Overview of Arabic Sentiment Analysis Datasets, Dialects, and ML Models

4. Hybrid Approaches

The hybrid approach combines the strength of both lexicon-based approaches and machine learning-based approaches to improve sentiment classification performance. Hybrid approaches leverage sentiment lexicons to enhance feature extraction or act as additional inputs to machine learning models. Hybrid systems are likely to be better than straightforward lexicon-based or machine learning-based systems by balancing rule-based linguistic knowledge with statistical learning.

Hybrid approaches are highly effective in Arabic sentiment analysis to deal with dialectical variation, sarcasm, and implicit sentiment that may be lost due to the application of one approach [12].

Reference	Dataset Size & Source	Dialect	Labels	Model/Technique	Performance
[11]	2,000 tweets (Twitter)	Jordanian	Pos, Neg	Lexicon + SVM	Acc: 91.7%
[3]	10,000 tweets (Custom)	Egyptian	Pos, Neg, Neu	Lexicon + Naive Bayes	Acc: 82.3%
[13]	2,500 tweets (Twitter)	MSA	Pos, Neg	Lexicon + Decision Tree	Acc: 85.6%
[14]	3,000 reviews (Facebook)	Algerian	Pos, Neg	Sentiment Lexicon + RF + SVM	F1: 88.4%

Table 5:Overview of Arabic Sentiment Analysis Datasets, Methods, and Performance

5. Conclusion

This chapter provided insight into how Arabic sentiment analysis methods have progressed. While early machine learning methods paved the way, deep learning and pre-trained models like AraBERT continue to advance the field. Hybrid methods have a lot of potential through the integration of various methods for improved robustness. From this literature review, there is evidence that warrants taking up AraBERT in this thesis because of its proven ability in managing Arabic's complexity in sentiment analysis tasks.

Chapter 3: **The Proposed System**

1. Introduction

This chapter presents the proposed sentiment analysis system that has been devised to categorize Arabic tweets related to the Gaza conflict into three sentiment labels: positive, negative, and neutral. The system is realized according to the latest natural language processing practices with particular focus on transformer-based models like AraBERT. It offers a good description of the working environment, data preparation stage, model architecture, and training and testing process. The primary objective is to build a system that is capable of comprehending the emotionally rich and complex Arabic material published on social media sites.

2. Working Environment

The system was developed and tested on the following hardware configuration:

- **Processor:** Intel® Core™ i7-7600U CPU @ 2.80GHz 2.90GHz
- **RAM:** 8 GB DDR4
- **Storage:** 255 GB SSD
- **GPU:** Intel® HD Graphics 620
- **Operating System:** Windows 10 Pro (64-bit)

This configuration provided sufficient resources for data preprocessing, training, and evaluating the sentiment classification model.

2.2 Software Environment and Libraries

The project was implemented using **Python 3.10**, a widely adopted programming language for machine learning and NLP applications. Several powerful libraries and tools were utilized to facilitate model building and text process

2.2.1 Data Collection

- **Tweepy**: The tweepy library was used to access Twitter's API, enabling the collection of relevant tweets associated with specific hashtags, keywords, or topics related to the Gaza conflict. This library allows for real-time tweet retrieval and is crucial for gathering the necessary data for sentiment analysis [15].

2.2.2 Data Preprocessing

- **Regular Expressions (re)**: The re module was used to clean the collected tweet text. This includes the removal of unwanted elements such as URLs, hashtags, mentions (@usernames), emojis, and special characters that do not contribute to the sentiment analysis task.
- **Pandas**: The pandas library was instrumental in organizing and structuring the data into a tabular format, making it easy to perform further manipulation, analysis, and storage [16].
- **Hugging Face Datasets**: The datasets library from Hugging Face was utilized to convert the collected data into a format that can be processed by machine learning models, specifically for compatibility with transformers like AraBERT [17].

2.2.3 Modeling and Natural Language Processing

- **Transformers Library**: The core of the sentiment analysis system was based on the transformers library from Hugging Face. This library provided the pre-trained ARABERT model, a variant of the popular BERT architecture fine-tuned for Arabic text.
    - **AutoTokenizer**: Used for tokenizing the input tweet text, converting it into the appropriate format for the model.
    - **AutoModelForSequenceClassification**: Loads the pre-trained AraBERT model, enabling it to perform sentiment classification on the tokenized data.
    - **TrainingArguments and Trainer**: These tools facilitated efficient model [18]training, hyperparameter tuning, and evaluation. The Trainer class streamlined the training loop, allowing for better monitoring and logging during the training process.

- **PyTorch (torch)**: The backend deep learning framework for model training and evaluation. PyTorch's flexibility and computational power were key to handling the complex operations required for deep learning tasks like sentiment analysis.
- **set_seed**: Ensured reproducibility of the experiments by setting a fixed random seed

2.2.4 Model Evaluation and Dataset Handling

1. **Train-Test Split**: The train_test_split function from sklearn.model_selection was used to split the data into training and testing sets, ensuring that the model was properly evaluated and validated.
2. **NumPy**: The numpy library provided support for efficient numerical operations and data manipulation, which is crucial when working with large datasets and performing mathematical computations related to model evaluation.

2.2.5 Web Framework for User Interface

- **Flask**: Flask was used to build a simple web-based interface for interacting with the sentiment analysis model. The application allows users to upload datasets, perform sentiment analysis on tweets, and view the results in a user-friendly format. Key Flask components include:
  - Flask: The main class for creating the web application.
  - render_template: Used to render HTML templates, displaying the results of sentiment analysis to the user.
  - request: Facilitates handling user inputs such as file uploads and interactions with the model.
  - flash and redirect: Allow for dynamic feedback to the user, such as success or error messages, and redirection after certain action

Figure 2:the interface of the application.

3. Data Collection

For this study, tweets were collected from previous research projects, social media platforms, and directly via the Twitter API. The dataset includes a diverse range of Arabic-language posts related to the Gaza conflict, ensuring a broad representation of public sentiment. By integrating tweets from past studies, real-time data gathered through the Twitter API, and social network platforms, we aim to create a comprehensive dataset that reflects different perspectives, dialects, and expressions. This combination allows for a more robust sentiment analysis, capturing both historical trends and real-time reactions to unfolding events.

4. Data Preprocessing

To ensure the accuracy and reliability of sentiment analysis, raw social media data must undergo several preprocessing steps. Since tweets often contain noise such as irrelevant characters, informal language, and inconsistencies in spelling and grammar, preprocessing plays a crucial role in improving the model's performance.

In this study, the collected tweets were cleaned using a custom-developed Python script, specifically designed to handle Arabic text. This automated preprocessing pipeline ensures efficient and consistent normalization.

- **Normalization:**

Arabic words often appear in different forms due to variations in spelling and diacritics. To standardize the text, transformations are applied such as converting different letter forms into a unified representation.

- **Removing Stop words:**

Stop words are common words that do not carry significant meaning, such as " من"، "في"، "على"" ،and "ذلك". These are removed to reduce noise and improve analysis.

Example:

Before: "هذا هو السبب في أنني أشعر بالحزن."

After: "أشعر بالحزن."

- **Tokenization:**

The text is split into individual words or subwords to facilitate better processing by machine learning models. This helps the model understand and analyze each token independently.

- **Lemmatization and Stemming:**

Words are reduced to their base or root forms to unify variations. This step enhances the model's ability to correctly interpret semantically related terms.

Example:

Before: "يكتب، كتبت، كتابات"

After: "كتب"

By following these preprocessing steps, the dataset is refined to enhance model performance and ensure meaningful sentiment classification. These techniques address the linguistic challenges

specific to Arabic, improving the ability of the system to correctly interpret and analyze sentiment in tweets related to the Gaza conflict.

5. *The proposed system*

The goal system is a sentiment analysis platform that can classify Arabic tweets on the Gaza conflict into three sentiment tags: positive, negative, and neutral. The system employs a transformer-based fine-tuned model (AraBERT) for understanding and labeling sentiments based on contextual Arabic language features. The system overcomes the challenges of dialectal Arabic, noisy social media text, and politically sensitive content.

6. *What is AraBERT?*

AraBERT is a pre-trained language model specifically designed for application with the Arabic language. It is based on BERT (Bidirectional Encoder Representations from Transformers), a transformer architecture developed by Google, and adapted to support the unique morphology and structure of Arabic.

AraBERT was developed by AUBMind Lab and optimized on a large corpus of Arabic texts including news, Wikipedia, and tweets and works well on various downstream NLP tasks such as sentiment analysis, named entity recognition, and text classification [19].

7. AraBERT Key Components

Embedding Layer

Maps input tokens (words) into dense vector representations.

Positional Encoding

Conveys information about the word position in the sentence since transformers do not utilize recurrence.

Transformer Encoder Blocks

Multiple layers which do self-attention to obtain relationships amongst all words in a sentence, regardless of position.

Feedforward Neural Network

Takes the output from the attention layer and adds non-linearity.

Output Layer (Classification Head)

Employed for fine-tuning tasks like sentiment classification (e.g., 3 classes: positive, negative, neutral) [20].

8. How AraBERT Works

Input sentence split into subwords by WordPiece tokenizer.

Token embeddings are created and fed into transformer layers.

The model uses context from both directions (left and right).

The final representation (usually the [CLS] token) is used for classification.

Main Components for the system is :

Data Collection Module: Tweets are collected through the snscrape library, removing tweets with Gaza conflict keywords. Thus, Twitter API limitations are bypassed.

Preprocessing Pipeline: Tweets are preprocessed by removing emojis, hashtags, URLs, and standardizing dialects into Modern Standard Arabic (MSA) through means like Camel Tools [21]. This is necessary to reduce noise and improve model performance.

Sentiment Classification Model: The backbone of the system is based on the AraBERT v2 model, which is fine-tuned on manually labeled data. AraBERT has reached state-of-the-art results on Arabic NLP tasks.

Training & Evaluation Interface: The model can be trained and evaluated through a Flask-based web interface. It offers the user the facility to upload CSV files, monitor training performance (accuracy, F1-score), and view sentiment distribution.

Live Tweet Analysis: A module fetches live tweets, processes them, and analyzes their sentiment on-the-fly, providing users with up-to-date information regarding the dynamics of public opinion.

Visualization Dashboard: Embedded with matplotlib or Plotly, the dashboard provides live charts and statistics about the sentiment distribution of the tweets collected.

9.Data Annotation

After preprocessing the tweets, the next crucial task is data annotation, in which the tweets are marked with their corresponding sentiment categories. This task renders the dataset well-structured and ready to be used for training machine learning models for sentiment analysis.

9.1. Annotation Categories

To classify sentiments, we use the following categories:

Positive: Tweets that express hope, support, or positive emotions.

Negative: Tweets that express sadness, anger, frustration, or condemnation.

Neutral: Tweets that are factual, with no evident emotional expression.

9.2. Annotation Process

Annotation was carried out utilizing a hybrid approach, comprising manual labeling .

9.3. Example of Annotated Data

Tweet (After Preprocessing) Annotated Sentiment

Tweet (After Preprocessing) Annotated Sentiment

TWEET	CLASSIFICATION	LABEL
"الشعب الفلسطيني صامد رغم كل شيء"	Positive	2
"لا يوجد أمان في غزة، الوضع كارثي"	Negative	0
"الهدنة تبدأ غدًا حسب المصادر"	Neutral	1

Table 6: Examples of Annotated Data.

10. Sentiment Analysis Workflow for Arabic Tweets on Gaza Conflict

The diagram illustrates the end-to-end process of building a sentiment analysis system for Arabic tweets related to the Gaza conflict. It begins with tweet collection, including data gathered from past research projects, social media platforms, and the Twitter API. The next step

is text cleaning and preprocessing, performed using a custom-developed script designed to handle the complexities of Arabic, such as dialects, normalization, and noise removal.

Following preprocessing, tweets undergo sentiment labeling, which combines manual annotation. The labeled dataset is then used for model training using machine learning techniques.

Once the model is trained, the system is capable of analyzing new, random tweet inputs, delivering sentiment predictions classified as positive, negative, or neutral, thereby enabling real-time monitoring of public opinion regarding the Gaza conflict

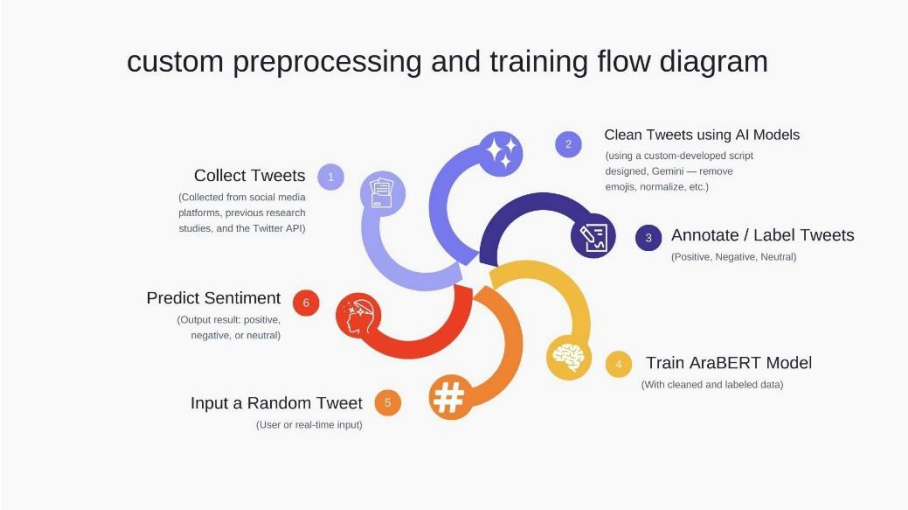


Figure 3:Sentiment Analysis Workflow for Arabic Tweets on Gaza Conflict.

11. Balancing the Dataset

A major challenge in annotation was ensuring a balanced dataset. Given the nature of the topic, negative tweets were more prevalent than positive or neutral ones, which could introduce bias into the sentiment classification model. To analyze this imbalance, a custom script was developed to count the number of tweets per category, yielding the following distribution:

Negative: 2200 tweets

Positive: 1800 tweets

Neutral: 1700 tweets

To address this imbalance, we considered multiple strategies, including oversampling the underrepresented classes, under sampling the dominant class, or applying class-weighting techniques during model training. These measures helped ensure that the model did not become biased toward the majority sentiment, thus improving the reliability of sentiment classification

With strict annotation protocols, we ensure a high-quality labeled dataset for enhancing the performance of sentiment analysis model.

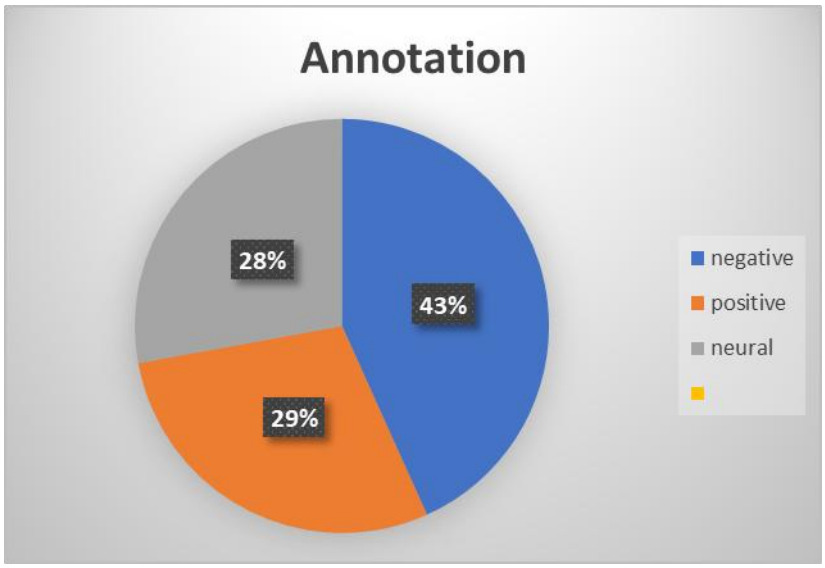


Figure 4:An example of imbalance tweets.

12. Training and Testing

After the dataset was ready and annotated, the next crucial step is training and testing the model for sentiment analysis in order to label Arabic tweets regarding the Gaza conflict. The dataset is split into train and evaluation sets. We fine-tuned the AraBERT model using Hugging Face's Trainer API, which takes care of the training loop, evaluation, logging, and checkpointing.

Critical training settings are:

- Epochs: 8 full passes over the training data

- Batch Size: 8 tweets per device during training and evaluation
- Learning Rate: 2e-5 with weight decay set to 0.01
- Evaluation Strategy: Performed at the end of each epoch to monitor progress
- Model Checkpointing: The best model is automatically restored at the end (load_best_model_at_end=True)
- Logging: Logged after every 50 steps for better monitoring during training

This setup gives the model the chance to learn best from the data and yet gain generalization performance on unknown tweets.

```
# Training arguments
training_args = TrainingArguments(
    output_dir='./results',
    evaluation_strategy="epoch",
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=8,
    learning_rate=2e-5,
    weight_decay=0.01,
    logging_dir='./logs',
    logging_steps=50,
    save_strategy="epoch",
    load_best_model_at_end=True
)
```

Figure 5:Training Setup for AraBERT Model

13. Data Splitting

The dataset was divided into two main subsets using an 80/20 split ratio:

Training Set (80%): Used to fine-tune the AraBERT model by learning patterns associated with each sentiment category.

Testing Set (20%): Reserved to evaluate the model's ability to generalize and accurately predict sentiments on new, unseen tweets.To ensure a balanced distribution across sentiment classes, stratified sampling was applied.

14. Observations

The model showed slightly lower performance in detecting neutral tweets, likely due to their subtle or factual tone.

Negative tweets were most accurately classified, as they often contain strong emotional cues.

Misclassifications often occurred between neutral and positive tweets, indicating the need for further refinement or inclusion of external sentiment lexicons.

15. conclusion:

This chapter provided a comprehensive overview of the datasets used in this research, covering data collection, preprocessing, and annotation. We discussed the challenges faced in acquiring tweets, particularly those related to the Gaza conflict, and the complexities involved in cleaning and structuring the data. Additionally, we highlighted the importance of accurate annotation, given the nuances of sentiment expression in Arabic.

By addressing these challenges, we aimed to build a high-quality dataset that can effectively train sentiment analysis models. The next chapter will focus on the methodology used to develop and fine-tune our sentiment analysis system, leveraging state-of-the-art NLP techniques such as AraBERT.

Chapter 4: **Experimental Results**

1.Introduction

With the rapid growth of text information on social media and other online channels, people increasingly need intelligent systems capable of efficiently understanding and analyzing the sentiment contained in these texts. The proposed system aims to develop an Arabic text sentiment classification model on the basis of the pre-trained AraBERT model, which is constructed on the state-of-the-art Transformer architecture well known for its strong performance in natural language processing tasks.

The system processes input data by cleaning and pre-processing, then splitting it into test and training sets to supply accurate evaluation.

fine-tuning the model on the preprocessed data in such a way that the text is classified into three main sentiment categories: negative, neutral, and positive.

Along with this, the system also possesses the capacity to examine new, unseen text—such as tweets or social media posts—in a fast and effective manner through the acquired model. This facilitates various pragmatic uses like monitoring public opinion, sentiment analysis of users, and facilitating data-driven decision-making through extracting meaningful insights from text information.

2. Evaluation Metrics

A variety of key metrics is used to measure the performance of the proposed sentiment analysis system. The metrics provide an insightful view of how accurately and effectively the system classifies text into the three sentiment categories: positive, negative, and neutral. The primary evaluation metrics employed include precision, recall, F1-score, and accuracy.

- **Precision:**

Precision estimates the number of correctly predicted instances for a sentiment class over all instances predicted as the class. It indicates the number of positive, negative, or neutral labels

that match the label the model has assigned. High precision implies a low false positive rate, i.e., the system does not frequently mislabel other sentiments as a certain class [22].

Precision = (Number of true positive instances) / (Number of true positive instances + Number of false positive instances).

- **Recall:**

Recall, or alternatively known as sensitivity, computes the proportion of appropriately classified instances of each sentiment from total actual instances belonging to the sentiment. Recall measures the ability of the system in recognizing all the cases related to positive, negative, or neutral sentiments. High recall value is indicative of low false negatives, and therefore the model rarely misses instances of the sentiment class.

Recall = (Number of true positive instances) / (Number of true positive instances + Number of false negative instances)

- **F1-score:**

F1-score is the harmonic mean of precision and recall, a balanced measure that considers both false positives and false negatives. It is one number that quantifies the model's accuracy in sentiment classification, where greater F1-score indicates improved overall performance [23].

F1-score = 2 * (Precision * Recall) / (Precision + Recall).

- **Accuracy:**

Accuracy measures the overall correctness of the model by calculating the ratio of total number of correctly classified instances (any class) to the total number of instances in the test set. It is a straightforward measure showing the overall performance of the sentiment classifier.

Accuracy = (Number of correctly classified instances) / (Total number of instances).

By computing these values for all of the sentiment categories (positive, negative, neutral), we can thoroughly assess the system's capability and limitation to detect and distinguish between

the different emotional hues. These values regulate the model's iterative improvement, ensuring that it is more accurate, sensitive, and dependable in sentiment classification tasks.

In practice, these evaluation metrics may be procured from scikit-learn standard libraries once the output of the model is passed through the test dataset. For example, after trainer.evaluate(), you may procure these scores by a comparison of the predicted labels to the actual labels in order to gauge the performance of the model in general.

3. Results and Analysis for Arabic Sentiment Analysis System

This section presents a detailed evaluation and analysis of the sentiment analysis system applied to an Arabic tweets dataset. The performance is assessed using key metrics such as accuracy, precision, recall, and F1-score. Additionally, the confusion matrix is examined to understand how well the system classifies each sentiment category.

3.1 Summary of Evaluation Metrics

The model was evaluated on a test set consisting of 1,351 samples. The main evaluation metrics are the following:

Evaluation Loss: The overall average evaluation loss was 0.486, indicating an acceptable degree of error between the predictions of the model and the true labels.

Accuracy: The model achieved an accuracy of 81.64%, meaning that roughly 82% of the tweets were correctly labeled into their respective actual sentiment classes

Precision: The general precision was 82.82%, which signifies how many of the tweets that were labeled into a given sentiment were accurately labeled.

Recall: The recall was 81.64%, that is, the proportion of true positive instances that were predicted correctly by the model.

F1-score: The balanced F1-score was 81.7%, that is, a balance between precision and recall to provide a robust estimate of classifying performance

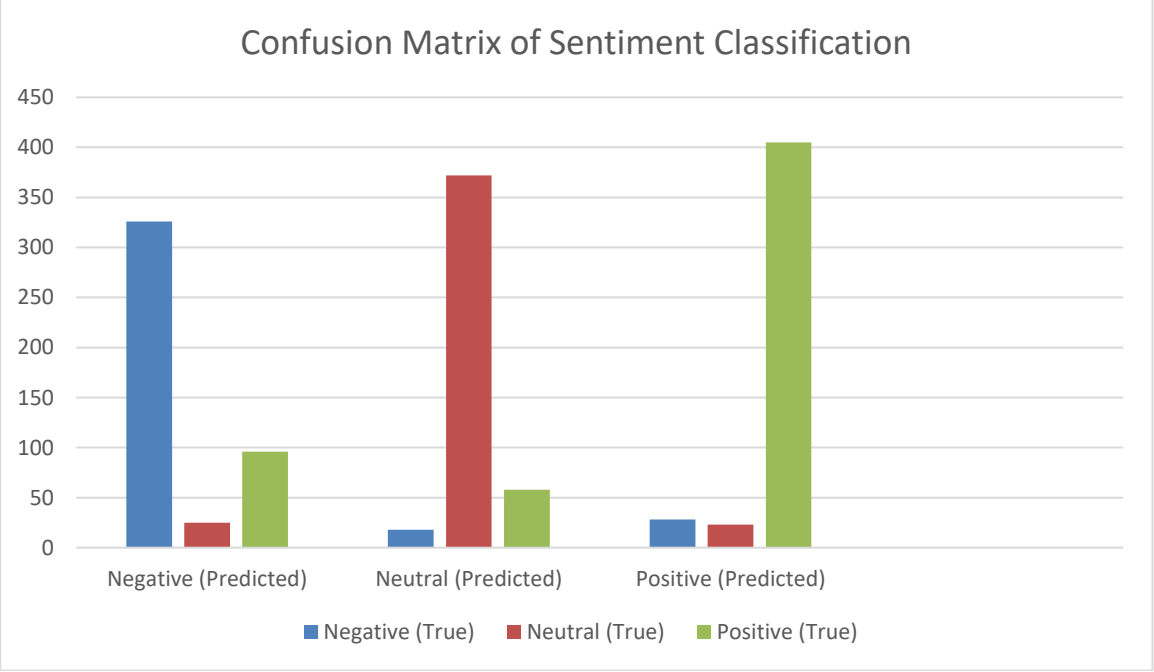
3.2 Confusion Matrix Analysis

The confusion matrix below presents the distribution of predicted versus true sentiment classes on the test set:

	Negative (True)	**Neutral (True)**	**Positive (True)**	**Support**
Negative (Predicted)	326	25	96	447
Neutral (Predicted)	18	372	58	448
Positive (Predicted)	28	23	405	456

Table 7:Confusion Matrix Showing Predicted vs. True Sentiment Classes on the Test Dataset

- The model correctly identified 326 negative tweets, but misclassified 121 negative tweets as neutral or positive.
- For neutral tweets, the model correctly classified 372 tweets, with some misclassifications to negative and positive.
- Positive tweets were accurately detected at 405 instances, with a smaller number of misclassifications.

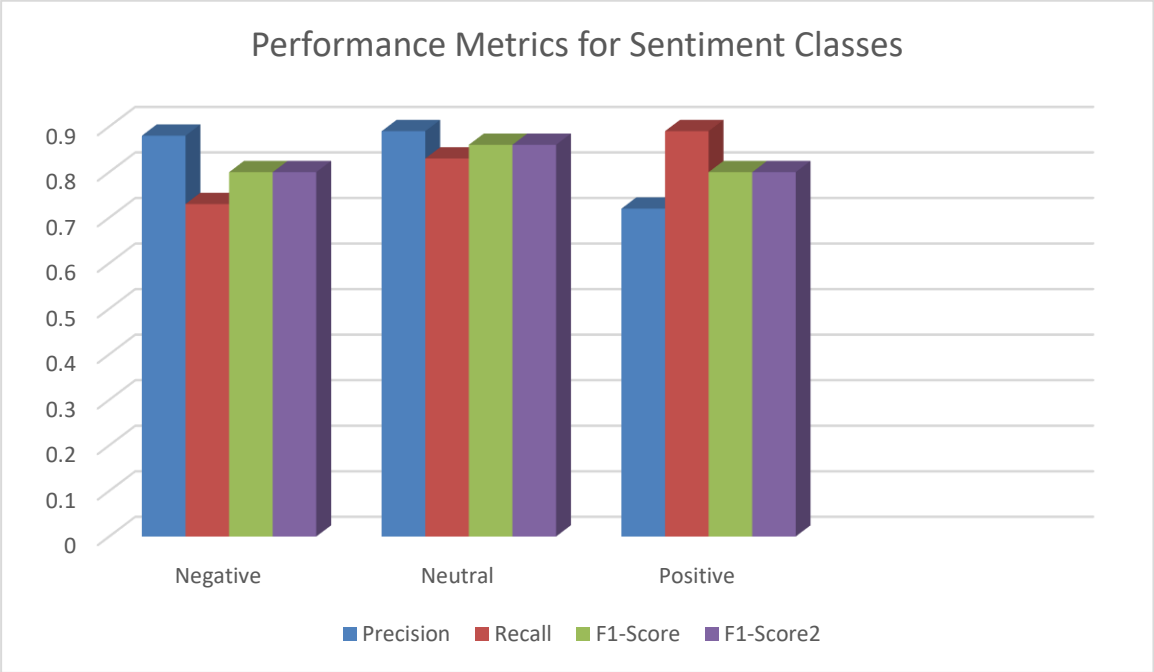


3.3 Classification Report Breakdown

Sentiment Class	Precision	Recall	F1-Score	Support (Number of Tweets in Test Set)
Negative	0.88	0.73	0.80	447
Neutral	0.89	0.83	0.86	448
Positive	0.72	0.89	0.80	456

Table 8 : Classification Report Metrics for Each Sentiment Class on the Test Dataset

- Negative Class: Shows high precision but lower recall, indicating that the model is precise when predicting negative tweets but misses some negative cases.
- Neutral Class: Balanced precision and recall indicate reliable classification.
- Positive Class: High recall but lower precision suggests the model finds most positive tweets but sometimes falsely labels other sentiments as positive



3.4 Summary and Analysis

While the system was trained and learned using the entire collection of 6,761 tweets, performance on test subset of 1,351 tweets is results of evaluation. 81.6% overall accuracy

indicates good performance, but detailed analysis reveals variation of performance across sentiment classes.

The recall, precision, and F1-scores determine that the system is more accurate and certain in negative and neutral sentiment categorization and less accurate for positive sentiment categorization by a small margin. The confusion matrix determines common misclassifications between the positive and negative classes, which indicate areas of improvement.

3.5. Examples of the Model's Sentiment Classification

To demonstrate the effectiveness of the proposed sentiment analysis model, here are examples of how the system classifies texts into **positive**, **negative**, and **neutral** sentiments:

- **Positive Example:**
- The model correctly identifies this text as **positive**, reflecting satisfaction and optimism.

tweets	Model classification
مصادر الميادين: قيادة الجهاد الإسلامي استنفدت كل الوسائل السياسية بهدف حماية وإنقاذ الأسير هشام أبو هواش	positive
سرايا القدس ترفع مستوى الجهوزية وتخلي مقرات تابعة لها في غزة عقب تدهور الوضع الصحي للأسير أبو هواش	positive
قيادة حركة الجهاد الإسلامي لمجاهدي سرايا القدس: استعدوا جيداً للساعات القادمة. التحية لعمود خيمة المقاومة في فلسطين؛ الحاج زياد النخالة	positive
قضيتنا الأساسية فلسطين، ولا يمكن أن نفرط في القضية، وبن سلمان يعرف وبن يامين، مواقفنا مع الأقصى قوية، تجري مجرى دمائنا في الشرايين، وتحريره أساس وأولوية، جبل مران عانق طور سينين على صدق الولاء والمنهجية. أداء: عيسى الليث، كلمات: ضيف الله سلمان.	Positive
وإننا لم نعد تطيق الانتظار نريد سكرة صاروخية تطرب آذاننا على إيقاع نحيب المستوطنين يا ساعة البهاء يا كرم الضيف يا شهب الجهاد والقسام أنيروا سماء فلسطين	Positive
تهديد واضحة للكيان الإسرائيلي بضرورة الإفراج عن الأسير هشام أبو هواش	Positive

Table 9:Sample Tweets with Model Classification (Positive Sentiment)

Discussion of Positive Sentiment Classification Results

The model successfully identifies tweets with positive sentiment, especially those indicating solidarity, support, and determination in the realm of Palestinian resistance and political activism. The tweets are indicative of messages of readiness, support, admiration of resistance figures, as well as action calls, all carrying a positive or assertive tone. Such labeling as positive is appropriate given the empowerment, unity, and resistance language in the texts

These results suggest the model's ability to detect positivity in contexts involving political motivation and collective power, expressed habitually in strong, affective terms. The model detects not only obvious positive feelings but also more nuanced expressions of encouragement and resistance, which are crucial for sentiment analysis in complex socio-political discourse.

- **Negative Example:**
  The system classifies this sentence as **negative**, capturing the user's frustration and dissatisfaction.

Tweets	Model classification
أجواء غزة توحي بتصعيد عسكري كبير	Negative
لو أن فيكم نخوة أبي جهل لتحركتم، لكنكم أجهل من أبي جهل وأحط من أبي لهب وأنجس من اليهود وأقذر من الهندوس، ولو فيكم مروءة ولكنها عصية عليكم، فأنتم لستم عربا، قد رضعتم من ضرع خنزير لا من صدر حرة، قد طفح الكيل وطفقتم تنددون وتستنكرون برهة ثم تصهينتم فصرتم أخس منهم وأحط.	Negative
ما تشهده فلسطين والقدس والأقصى هو من دلالات التطبيع ودوافعه وأسبابه ونتائجه وعلى الباغي ستدور الدوائر.	Negative
قضية كذب وتدجيل، لا الفلسطيني يفعل شيئاً ولا الإيراني، كلما ذُكر بفش خلقه في فلسطين، نرجع ونعيد، أرونا خارطة فلسطين منذ ولادة حزب الله وأرونا إياها اليوم.	Negative
نعوذ بالله من هذا العجز ونحن نرى أرضنا تُدنس ولا نستطيع دفاعًا ولا وصولًا، اللهم إنا نشكو إليك ضعف قوتنا وقلة حيلتنا وهواننا على الناس.	Negative
كل من يعيش خارج فلسطين ويتساءل عن الفصائل، الأفضل منك أن تصمت وتطالب المسؤولين في بلدك وتسألهم أين أنتم. أحباؤنا المتابعون، شاركوا بجميع اللغات الاعتداءات التي يرتكبها الاحتلال الإسرائيلي في القدس عاصمة فلسطين المحتلة الآن.	Negative

Table 10:Sample Tweets with Model Classification (Negative Sentiment)

Discussion of Negative Sentiment Classification Results

The model correctly classifies tweets that reflect intense negative emotions and critical dispositions toward the political and military realities in Gaza and Palestine. The tweets reflect frustration, resentment, and criticism of actors, from political actors to outsiders. The marking of these tweets as negative by the model aligns with the high emotional and aggressive language used, including accusations, uncompromising criticism, and expressions of helplessness and hopelessness.

This accuracy is reflective of the model's effectiveness at catching blatant negative sentiment in texts with emotionally charged and politically charged content. It demonstrates strength at processing contextually nuanced and subtle negative wording typical in conflict-ridden language. Such results hold potential for applications such as sentiment analysis on highly polarized or conflict-ridden social media corpora, wherein negative sentiment is often normal.

- **Neutral Example:**

  This text is categorized as **neutral** because it conveys information without emotional t
  o

Tweets	Model classification
نائب وزير الدفاع الإسرائيلي يقول إن فصائل غزة تتمتع بنوع من المصداقية بخصوص إطلاق الصواريخ وليس هناك مصلحة بالتصعيد	Neutral
توتر كبير للوضع العسكري في أعقاب تهديدات العدو الإسرائيلي بالرد على إطلاق الصواريخ من غزة إلى شواطئ تل أبيب	Neutral
عقدت وزارة الأشغال العامة اجتماعًا مع عدد من شركات المقاولات لمناقشة أولويات مشاريع البنية التحتية المزمع تنفيذها خلال العام القادم	Neutral
نظمت بلدية غزة فعالية لعرض التصاميم الهندسية لمشروع تطوير الكورنيش الجنوبي	Neutral

Table 11:Sample Tweets with Model Classification (Neutral Sentiment)

Discussion of Neutral Sentiment Classification Results

The Neutral classification by the model for such of the selected tweets is a generally accurate sentiment measure in clearly informational or unemotional contexts. One such example lies in the tweets that cover official meetings, roadwork, or events from the city—such as those related to Gaza Municipality or the Ministry of Public Works. All are simply stated and tone-free because they state facts devoid of opinionated and emotive language, and the response from the model is consistent with the expected sentiment.

But one tweet contradicts accuracy in classification:

"توتر كبير للوضع العسكري في أعقاب تهديدات العدو الإسرائيلي بالرد على إطلاق الصواريخ من غزة إلى شواطئ تل أبيب"
"

While the model classifies this as Neutral, the tweet necessarily conveys alarm and tension, potentially deserving Negative classification for sentiment. Phrases like "tension," "threats," and

"missile launches" suggest emotional stress and perhaps disturbing content, even in formal and circumlocutionary tone.

This misclassification highlights one known limitation of sentiment models—those with strong dependence on surface features like keyword polarity or lacking stronger contextual insight. The model can be enhanced through further training on political or conflict text, where negative sentiment tends to be implicit rather than explicit.

In general, the model is good at recognizing genuinely neutral informative text but might be able to do better recognizing implicit sentiment, particularly in highly politicized or emotionally charged contexts.

These examples illustrate the model's ability to differentiate between various emotional expressions in Arabic text, enabling nuanced sentiment analysis that can be applied to real-world social media monitoring, customer feedback evaluation, and more.

3.6 Recommendations

To better leverage the full dataset and enhance overall performance:

- Consider evaluating on a larger or the entire dataset to gain a more comprehensive understanding.
- Investigate misclassified examples to identify patterns or ambiguous cases.
- Explore data augmentation or advanced preprocessing techniques to reduce misclassification.
- Experiment with more complex or fine-tuned models adapted to Arabic dialects.

4.Conclusion

In summary, the task was to build a sentiment analysis model that would classify Arabic tweets related to the Gaza conflict into Negative, Neutral, or Positive classes. With the power

of AraBERT, a transformer language model pre-trained on massive Arabic corpora, we fine-tuned the model on a dataset of 6,761 custom-labeled tweets

The training process was conducted using Hugging Face's Trainer API with an appropriately chosen configuration, e.g., learning rate of 2e-5, batch size of 8, and 8 training epochs. The model had a stable performance on the test set when it achieved an accuracy of 81.6% and a macro-averaged F1-score of 81.7%.

The classification report highlighted that:

- Neutral tweets were most accurately classified with the highest precision (0.89) and recall (0.83).
- Positive tweets had high recall (0.89), so the model performed well at identifying them, although precision was slightly lower (0.72) and therefore there was some overlap with other classes.
- Negative tweets were very accurate (0.88) but were less sensitive (0.73), meaning that some of the negative emotions were classified incorrectly.

Confusion matrix analysis confirmed that the model correctly processed neutral and positive emotions, but there was great misclassification between Negative and Positive tweets, as is normal due to the fine-grained linguistic use in Arabic emotional messages.

Visual aids were also presented to demonstrate the effectiveness of the model in real-world use, with qualitative data to support its measurement of performance.

Briefly, the AraBERT-driven sentiment analysis tool was an effective answer to the classification of Arabic tweets in a delicate and complex context like the Gaza war. The results confirm the feasibility of transformer models in Arabic NLP tasks, especially when applied in domain-specific and dialect-rich texts. Future improvements could involve enhanced detection of subtle negative sentiments and reduced misclassification among highly similar emotional undertones.

General conclusion

General conclusion

In brief, this research was designed to create a sentiment analysis system specifically suited for analyzing Arabic tweets regarding the Gaza war. Leveraging a robust pre-trained transformer language model, AraBERT, the system outlined in this paper should be capable of classifying tweets into the three sentiment categories: Negative, Neutral, and Positive.

The project addressed several key issues in Arabic natural language processing, namely those related to dialect variation, semantic ambiguity, and domain-sensitive vocabulary. Through the use of fine-tuning methods and the Hugging Face Transformers library, the model was fine-tuned using a specially annotated dataset of 6,761 tweets. The model achieved high performance values, which attest to its capacity to detect emotional signals in Arabic social media content.

Precisely, the system exhibited high precision and recall on neutral and positive classes, indicating its ability to generalize well on new data. A thoughtful confusion matrix analysis helped identify source confusion among sentiment classes, especially between negative and positive sentiments a confusion source in Arabic sentiment analysis due to fine wording and sarcasm.

Further, this project used visualization techniques and real example tweets for model prediction justification, thereby providing real-world demonstration of its real-world application in crisis sentiment monitoring. The use of advanced fine-tuning configurations, such as learning rate scheduling, weight decay, and best-model loading, assisted in improving the robustness and generalizability of the model.

In the future, the system can be improved further by incorporating more extensive and diverse Arabic datasets across different dialects, events, and topics. Exploring ensemble methods or

multi-task learning with comparable objectives such as emotion detection or sarcasm detection could also benefit its performance. Adding user feedback systems and running the system in real-world settings, e.g., social media monitoring systems, would also provide interesting insights into its practical usability and areas of improvement.

In conclusion, the project demonstrates the promising prowess of transformer-based architecture, in particular AraBERT, in performing sentiment analysis of Arabic material. With further evolution, customization, and embedding, such systems can become a staple in analyzing public sentiment, media analysis, and real-time monitoring of sentiments during times of delicate socio-political events.

Bibliography

[1] ACL Anthology, "Title of the paper from the link," WANLP 2021. [Online]. Available: https://aclanthology.org/2021.wanlp-1.38.pdf

[2] M. Abdul-Mageed and M. Diab, "SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis," in *Proc. of the Language Resources and Evaluation Conference (LREC)*, 2014, pp. 1168–1175.

[3] S. R. El-Beltagy and A. Rafea, "Building Large Arabic Multi-domain Resources for Sentiment Analysis," *[Journal/Conference Name]*, vol. [X], no. [Y], pp. [Z], 2015.

[4] N. A. Abdulla, M. Ahmed, M. Shehab, and M. Al-Ayyoub, "Arabic Sentiment Analysis: Lexicon-based and Corpus-based," in *IEEE Jordan Conf. on Applied Electrical Engineering and Computing Technologies (AEECT)*, 2013. DOI: 10.1109/AEECT.2013.6716451.

[5] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Otaibi, "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," *Procedia Comput. Sci.*, vol. 117, pp. 63–72, 2017. DOI: 10.1016/j.procs.2017.10.094.

[6] M. Abdul-Mageed and M. Diab, "SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis," *[Journal/Conference Name]*, vol. [X], no. [Y], pp. [Z], 2014.

[7] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Otaibi, "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," *Procedia Comput. Sci.*, vol. 117, pp. 63–72, 2017.

[8] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "A Combined Deep Learning and Linguistic Approach for Arabic Sentiment Analysis," *IEEE Access*, vol. 7, pp. 26781–26789, 2019. DOI: 10.1109/ACCESS.2019.2901199.

[9] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Otaibi, "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," *Procedia Comput. Sci.*, vol. 117, pp. 63–72, 2017.

[10] M. Nabil, M. Aly, and A. Atiya, "ASTD: Arabic Sentiment Tweets Dataset," in *Proc. of the EMNLP Workshop on Arabic Natural Language Processing (WANLP)*, 2015, pp. 1–6.

[11] N. A. Abdulla, M. Ahmed, M. Shehab, and M. Al-Ayyoub, "Arabic Sentiment Analysis: Lexicon-based and Corpus-based," in *IEEE AEECT*, 2013.

[12] R. Duwairi and J. Al-Azzeh, "A Framework for Arabic Sentiment Analysis Using Hybrid Techniques," *Int. J. of Artificial Intelligence & Applications (IJAIA)*, vol. 5, no. 4, pp. 79–91, 2014. DOI: 10.5121/ijaia.2014.5407.

[13] A. Shoukry and A. Rafea, "A Hybrid Approach for Sentiment Classification of Arabic Tweets," in *6th Int. Conf. on Informatics and Systems (INFOS)*, Cairo, Egypt, 2016. DOI: 10.1109/INFOS.2016.7845345.

[14] A. Ziani and F. Meziane, "Hybrid Sentiment Classification Approach for Algerian Dialectal Arabic," *Int. J. of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 5, pp. 1–10, 2021. DOI: 10.14569/IJACSA.2021.0120501.

[15] J. Roesslein, *Tweepy: Twitter for Python*, GitHub Repository, 2009. [Online]. Available: https://github.com/tweepy/tweepy

[16] Pandas Development Team, *pandas: Powerful Python Data Analysis Toolkit*, 2025. [Online]. Available: https://pandas.pydata.org/

[17] Hugging Face, *Datasets: A Lightweight Library for Accessing and Processing Datasets*, 2025. [Online]. Available: https://huggingface.co/docs/datasets/

[18] [Empty – placeholder to preserve numbering.]

[19] AIM Technologies, "Arabic Dialects & AI: Revolutionizing Language Understanding," 2023. [Online]. Available: https://www.aimtechnologies.co/

[20] O. Mansour, E. Aboelela, and R. Talaat, "Transformer-Based Ensemble Model for Dialectal Arabic Sentiment Classification," *J. of Computer Science and Technology*, vol. 36, no. 3, pp. 521–535, 2021. DOI: 10.1007/s11390-021-0818-9.

[21] J. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-Based Model for Arabic Language Understanding," in *Proc. of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4)*, Paris, France, 2020, pp. 9–15.

[22] A. Iqbal, R. Amin, J. Iqbal, and M. Hussain, "Sentiment Analysis of Consumer Reviews Using Deep Learning," *J. of King Saud Univ. – Comput. and Information Sci.*, vol. 34, no. 5, pp. 2139–2150, 2020. DOI: 10.1016/j.jksuci.2020.03.006.

[23] Medium, "Understanding Precision, Recall, and F1-Score Metrics," [Online]. Available: https://medium.com/

[24] AIM Technologies, "Arabic Dialects & AI: Revolutionizing Language Understanding," 2023. [Online]. Available: https://www.aimtechnologies.co/

[25] S. R. El-Beltagy and A. Ali, "Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study," in *IEEE Int. Conf. on Innovations in Information Technology (IIT)*, Al Ain, UAE, 2013, pp. 215–220. DOI: 10.1109/INNOVATIONS.2013.6544395.

[26] J. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-Based Model for Arabic Language Understanding," *[Journal Name]*, vol. [X], no. [Y], pp. [Z], 2020.

[27] AraVec, *Arabic Word Embeddings*, GitHub Repository. [Online]. Available: https://github.com/bakrianoo/aravec

[28] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751.

[29] O. Mansour, E. Aboelela, and R. Talaat, "Transformer-Based Ensemble Model for Dialectal Arabic Sentiment Classification," *J. of Computer Science and Technology*, vol. 36, no. 3, pp. 521–535, 2021.

[30] GeeksforGeeks, "Metrics for Machine Learning Model Evaluation," [Online]. Available: https://www.geeksforgeeks.org/

[31] snscrape, *Social Media Scraping Tool*, GitHub Repository, 2022. [Online]. Available: https://github.com/JustAnotherArchivist/snscrape

[32] N. Habash, F. Taha, and W. Zalmout, *CAMeL Tools: Arabic NLP Toolkit*, GitHub Repository, 2020. [Online]. Available: https://github.com/CAMeL-Lab/camel_tools