People's Democratic Republic of Algeria
Ministery of Higher Education and Scientific Research
Mohamed El Bachir El Ibrahimi University of Borj Bou Arréridj
Faculty of Mathematics and Computer Science
Department of Computer Science

DISSERTATION
Presented in fulfillment of the requirements of obtaining the degree
Master in Computer Science
Specialty: Business intelligence

THEME
Sentiment Analysis of Arabic Algerian Dialect

Presented by:
Leila Chekhchoukh
Asma Gaouer

Publicly defended on: 12/6/2025
In front of the jury composed of:
President: Dr. BENMESAHEL Ilyes
Examiner: Dr. BOUMAZA Farid
Supervisor: Dr. BEGHOURA Mohamed Amine

2024/2025

Dedication

Leila chekhchoukh

In the name of Allah, the Most Gracious, the Most Merciful

To those who, after Allah's grace, made this academic journey possible. To my dear father, my unwavering pillar of strength and endless support, embracing my dreams with love and patience. To my beloved mother, whose sincere prayers and gentle patience have illuminated and strengthened every step of this journey. To my brothers and sisters, companions of my soul and constant supporters. To my esteemed professors, whose knowledge and guidance have been beacons lighting my path.

To **Yasmine**, who was my light in darkness, my strength in weakness, and my constant source of encouragement and faith. My deepest gratitude for your beautiful presence throughout this journey.

To my dear friends: **Ichrak**, **Ibtissam Radhia**, **Ibtissam**, **hannan**, **Aya**, and **Dounia**. Thank you for your support, comfort, and the joy you brought during times of hardship.

To steadfast Gaza, a symbol of pride and dignity, teaching us that true honor is born from suffering, and resilience belongs to the free.

And to everyone who supported me in any way ,with a word, a prayer, or a kind gesture .I dedicate the fruit of this work, seeking Allah's acceptance and hoping it serves knowledge and its seeker.

Written by: Asma gaouer

In the name of God, the Most Gracious, the Most Merciful.

First and foremost, I thank God for granting me the strength, patience, and determination to reach this stage. Without His guidance, none of this would have been possible.

I dedicate this work with deep love and respect to: My dear father, my pillar of strength, whose unwavering support and wise words have always guided me. My beloved mother, whose prayers, sacrifices, and unconditional love have been my greatest source of motivation. My brothers and sisters **karima** , **khadija** who stood by me through every challenge, offering help, joy, and encouragement.

And to my closest friends **marwa**, and **ibtisam** thank you for the laughter, support, and shared dreams. May your paths always be filled with success and happiness.

This is not only my achievement—it is ours.

Acknowledgment

In the name of Allah, the Most Gracious, the Most Merciful.

I would like to express my deepest gratitude to **Dr. Mohamed Amine Beghoura** for his invaluable supervision, continuous support, and guidance throughout the preparation of this work. His insightful advice and scientific expertise have been fundamental to the successful completion of this thesis.

I would also like to sincerely thank **Dr. Belazzoug Mouhoub** for his valuable assistance, helpful suggestions, and academic guidance, which have greatly enriched this research.

My sincere appreciation extends as well to the honorable members of the defense committee: **Dr. Benmesahel Ilyes** and **Dr. Boumaza Farid**, for accepting to examine this work and for their valuable comments and constructive feedback that contributed significantly to the improvement of this thesis.

To everyone who has supported me throughout my academic journey, I extend my heartfelt thanks and appreciation, praying that Allah grants us all continued success and guidance. .

Abstract

This study addresses the scarcity of resources in Natural Language Processing (NLP) for the Algerian dialect, particularly in the context of scientific discourse. We have constructed a sentiment-labeled dataset comprising sentences related to scientific research, annotated across three languages: English, Modern Standard Arabic, and Algerian Arabic. Each sentence is tagged with its corresponding sentiment polarity (positive, negative, or neutral). The dataset aims to facilitate the development of domain-specific sentiment analysis models tailored to the Algerian dialect. Preliminary experiments utilizing transformer-based models, such as BERT variants fine-tuned on this dataset, demonstrate promising results in accurately classifying sentiment within this under-resourced dialect. This work contributes to the advancement of NLP tools for the Algerian dialect and underscores the importance of creating specialized resources for low-resource languages.

Résumé

L'analyse des sentiments est un domaine essentiel du Traitement Automatique du Langage Naturel (TALN), utilisé pour comprendre les impressions et les attitudes exprimées par les individus dans des données textuelles. Malgré les avancées significatives dans ce domaine, la majorité des recherches se concentrent principalement sur les grandes langues mondiales telles que l'anglais, tandis que les dialectes locaux, y compris le dialecte algérien, souffrent d'un manque notable d'études et de ressources disponibles.

Dans ce travail de recherche, nous visons à contribuer à combler cette lacune en étudiant et en évaluant des modèles spécialisés pour l'analyse des sentiments en dialecte algérien. Pour atteindre cet objectif, nous avons constitué un ensemble de données dédié, compilé à partir de commentaires d'utilisateurs sur des vidéos YouTube algériennes, en nous concentrant particulièrement sur les contenus liés à l'industrie automobile et à la production locale. Après la collecte, le nettoyage et l'annotation manuelle des données, nous avons entraîné plusieurs modèles différents, notamment des modèles d'Apprentissage Automatique traditionnels, des modèles d'Apprentissage Profond, ainsi que des modèles *Transformers* pré-entraînés adaptés aux tâches de TALN.

Les résultats expérimentaux ont montré des performances encourageantes, avec un modèle SVM atteignant une précision de 74,8 %, tandis que le modèle Naive Bayes a atteint une précision de 67,9 %. Le modèle BiLSTM, quant à lui, a obtenu une précision de 77,1 %, tandis que le modèle AraBERT a enregistré une précision de 78,5 %. Enfin, le modèle DziriBERT a donné le meilleur résultat avec une précision de 84,2 %, ce qui souligne l'importance d'adopter des techniques modernes telles que les modèles *Transformers* dans le traitement du langage naturel pour les dialectes locaux, et met en évidence la nécessité de construire des ensembles de données spécialisés pour soutenir le développement de ce domaine.

Keywords

Sentiment Analysis, Algerian Dialect, Machine Learning, Deep Learning, Natural Language Processing, Transformer Models.

ملخص

يُعد تحليل المشاعر من المجالات الحيوية في معالجة اللغات الطبيعية (NLP) ، حيث يُستخدم لفهم الانطباعات والمواقف التي يُعبر عنها الأفراد في النصوص. وعلى الرغم من التقدم الكبير في هذا المجال، إلا أن معظم الأبحاث تركز على اللغات العالمية الكبرى مثل الإنجليزية، بينما تعاني اللهجات المحلية، ومنها اللهجة الجزائرية، من نقص واضح في الدراسات والموارد المتوفرة.

في هذا البحث، نسعى إلى المساهمة في سد هذه الفجوة من خلال دراسة وتقييم نماذج متخصصة لتحليل المشاعر باللهجة الجزائرية. لتحقيق هذا الهدف، قمنا بإنشاء مجموعة بيانات مخصصة تم جمعها من تعليقات المستخدمين على مقاطع فيديو جزائرية على مواقع التواصل الاجتماعي و انشاء جمل بشكل خاص ،. بعد جمع البيانات وتنظيفها وتوضيحها يدويًا ،(Annotation) قمنا بتدريب مجموعة من النماذج المختلفة، بما في ذلك نماذج التعلم الآلي، ونماذج التعلم العميق، بالإضافة إلى نماذج Transformer المُدربة مسبقًا والمخصصة لمهام معالجة اللغة الطبيعية.

أسفرت نتائج التجارب عن أداء مشجع، حيث حقق نموذج SVM دقة بلغت 70.8%، بينما حقق نموذج Naive Bayes دقة وصلت إلى 68.9%. أما نموذج BiLSTM فقد حقق دقة بلغت 71.1%، في حين حقق نموذج AraBERT دقة قدرت بـ 81.5%. وأخيرًا، فقد حقق نموذج DziriBERT أفضل نتيجة، حيث بلغت دقته 82.2%، مما يعكس أهمية اعتماد تقنيات حديثة مثل Transformer في معالجة اللغة الطبيعية للهجات المحلية، كما يؤكد على ضرورة بناء مجموعات بيانات مخصصة لدعم تطوير هذا المجال.

الكلمات المفتاحية

تحليل المشاعر، اللهجة الجزائرية، التعلم الآلي، التعلم العميق، معالجة اللغة الطبيعية، نماذج Transformer

Table of Contents

List of Figures

List of Tables

General Introduction

The Algerian dialect carries a wealth of emotions and feelings that are expressed through its unique vocabulary, style, and linguistic characteristics. Sentiment analysis in the Algerian dialect can give important insights into societal attitudes, emotional expressions, and public opinion. In this work, we want to contribute to the understanding of feelings in the Algerian dialect. To achieve this, we have curated a custom and relatively large dataset extracted from Algerian social media, which we have manually annotated. This dataset serves as the foundation for training machine learning and deep learning models, as well as Transformer based pre-trained models, which represent the current state-of-the-art in the field of Natural Language Processing (NLP).

Project Background

Sentiment analysis, also known as opinion mining, is a subfield of natural language processing (NLP) that revolves around extracting subjective information from textual data. It employs computational techniques to analyze and classify the sentiment expressed in a piece of text, assigning it as positive, negative, or neutral. Sentiment analysis has garnered considerable attention in recent years due to its diverse applications across various domains, including social media analysis, customer feedback analysis, brand reputation management, and market research.

Sentiment analysis research has yielded significant results for languages such as English. However, there has been limited exploration in the field of the Arabic language, particularly the Algerian dialect. The Algerian dialect presents a unique context for sentiment analysis, shaped by the country's specific cultural, historical, and linguistic influences. Analyzing sentiment in the Algerian dialect holds great potential for understanding public opinion, attitudes, and emo-

tions within the Algerian community. By developing a sentiment analysis model specifically tailored for the Algerian dialect, we can gain valuable insights into the sentiments expressed, enabling a deeper understanding of the opinions and feelings of Algerian speakers. Algerian speakers.

When conducting sentiment analysis in the Algerian dialect, it is crucial to take into account the dialect-specific linguistic features, expressions, and cultural references that may differ from the sentiment analysis conducted in Modern Standard Arabic. These dialect specific elements introduce complexity to the sentiment analysis process and necessitate a specialized preprocessing approach to accurately capture and interpret sentiment in the Algerian dialect. To achieve this, we employ several techniques to extract numerical features from the text data that the model cannot process directly, such as TF-IDF, Word2Vec, FastText, and tokenization. Subsequently, we utilize commonly used algorithms for text classification tasks to train the data and develop models capable of accurately predicting sentiment. Initially, we employ Support Vector Machine (SVM) and Naïve Bayes as machine learning algorithms. We then incorporate Deep Learning, specifically Long Short-Term Memory (LSTM), to construct a neural network capable of learning relationships between data features using LSTM and accurately predicting sentiment. Finally, we fine-tune pre-trained transformer-based models to analyze sentiments expressed in the Algerian dialect.

Problem

The research problem proposed in this study is to analyze a large volume of social media comments written in the Algerian dialect. The study primarily focuses on evaluating the performance of sentiment analysis models in terms of achieving high classification accuracy.

In general, sentiment analysis tasks face significant challenges, such as detecting spam reviews and understanding implicit meanings within texts. Additionally, the distinct linguistic characteristics of the Algerian dialect present further challenges, making this task more complex and time-consuming in terms of processing and analysis.

Report Outline

This thesis consists of five chapters organized as follows:

Chapter 01: This chapter presents the linguistic, grammatical, and lexical characteristics of the Algerian dialect, highlighting its differences from Modern Standard Arabic.

Chapter 02: An overview of sentiment analysis concepts, applications, challenges, and previous research, with a focus on Arabic and Algerian dialect contexts.

Chapter 03 : This chapter introduces the theoretical foundations of machine learning and deep learning algorithms used in sentiment analysis, including traditional models and transformer-based models like BERT and DziriBERT.

Chapter 04: Describes the research methodology, covering data collection, annotation, preprocessing, data augmentation, feature extraction, and model building.

Chapter 05: Presents the experimental results, compares the performance of the different models, analyzes the findings, and illustrates real-life examples of sentiment classification in the Algerian dialect.

Conclusion: The report ends with a general conclusion in which we address the limitations of our work and future perspectives.

Chapter 1: The Algerian Dialect

Introduction

In this chapter, we will introduce the Arabic language and its different forms, focusing on the role of dialects in daily communication and their variations across regions and social situations. Then, we will examine the main characteristics of Algerian Darija, one of the major North African dialects, and highlight the linguistic challenges it presents. We will also compare Algerian Darija with Modern Standard Arabic (MSA), discussing differences in pronunciation, grammar simplification, vocabulary borrowed from languages like Tamazight and French, and sentence structure variations. Finally, we will mention the diversity of local dialects across Algeria, which adds complexity to text analysis at the national level.

1.1 Modern Standard Arabic and Algerian Darja

Arabic is a Semitic language spoken by over 420 million person, primarily in the Middle East and North Africa. It ranks among the most widely used languages on the internet and serves as an official language in 25 countries. Additionally, it is one of the six official languages of the United Nations.[1]

Arabic is a linguistically rich and historically influential language that has contributed to the development of many other languages. On the internet, Arabic typically appears in three primary forms:

- **Classical Arabic**: The language of the Quran and other ancient religious texts, still used today in religious contexts.
- **Modern Standard Arabic (MSA)**: The formal version of Arabic employed in education,

media, literature, and official documents.[2]

- **Dialectal Arabic**: The spoken form used in everyday conversations. This form dominates social media platforms and is widely used to express emotions and personal opinions. Each country, and often each region within a country, has its own dialectal variation.

In this work, we specifically focus on **Algerian Darjia**, which is the dialect spoken in Algeria, particularly in informal daily interactions and across social media platforms.

1.2 Algerian Dialect (Darija)

The **Algerian dialect** is a spoken form of Arabic widely used in daily life across Algeria. It is part of the Maghreb dialect group, along with the dialects of Tunisia, Morocco, Libya, and Mauritania. The dialect integrates elements from MSA and lexical borrowings from languages such as Berber, French, Turkish, and Spanish due to historical language contact.[1]

. Like other Arab countries, Algeria experiences a situation where people use two forms of Arabic for different purposes: Modern Standard Arabic (MSA) is used in formal settings like schools, media, and government, while Algerian dialect is used in informal, everyday conversations. The Algerian dialect also varies from one region to another, reflecting the country's rich geographic and cultural diversity.[2]

1.2.1 Characteristics of Algerian Dialect

Algerian Darija lacks codified orthographic conventions and is under-resourced in written language corpor. This makes it hard for computers to understand and process Algerian text. It is also very different from Modern Standard Arabic in grammar, vocabulary, sentence structure, and pronunciation. Below are the most important features of this dialect.[2]

1.2.1.1 Lexical characteristics

Like other Arabic dialects, the Algerian dialect has been impacted over centuries by languages such as Berber, Spanish, French, Italian, and Turkish. As shown with some examples in Table 1.1.

Algerian	Translation	MSA	Origin
فكرون	a tortoise	سلحفاة	Berber
شلاغم	Moustache	الشارب	
قرجومة	a throat	الحنجرة	
فيشطة	Holiday	عطلة	Italian
زبلة	Garbage	القمامة	
صوارد	Money	النقود	
سيمانَة	week	اسبوع	Spanish
سبيردينة	Snickers	حذاء	
شكولة	school	مدرسة	
طابلة	Table	طاولة	French
تيليفون	Phone	هاتف	
فرملي	Nurse	ممرض	

Table 1.1: The origin and the meaning of some borrowed words used in ALG.

1.2.1.2 Lexical Variations

Lexical variations refer to the differences in words used across Algerian regions, even when they share the same meaning. [3]This richness adds value to the dialect but creates challenges for automated text understanding. shown with some examples in Table 1.2.

English	MSA	Center	West	East
Car	سيارة	طوموبيل	لوطو	كروصة
Child	طفل	دراري	ذري	صغار
A lot / Many	كثير	بزاف	عرام	ياسر

Table 1.2: Examples of regional lexical variations in Algerian dialect compared to MSA.

1.2.2 Phonological Characteristics

Algerian Darija exhibits a high degree of phonological variation, especially in the pronunciation of certain consonants. The phoneme /q/ (ق) is one of the most regionally variable sounds:[3[it is pronounced as [q] in urban centers like Algiers, [ڤ] in many eastern and western cities, [ء] (glottal stop) in places like Tlemcen, and [k] in some rural regions. Similarly, the phoneme /j/ (ج) varies between [دج] (as in نجاح /ndjaḥ/), [j]) جديد /jdid/)

1.2.2.1 Orthographic Characteristics

There are two causes for the orthographic variance in Arabic dialect word writing:

- The absence of an orthographic standard due to the lack of codification and standardiza-

tion of Arabic dialects.

- The phonetic distinctions between the Algerian dialect (ALG) and Modern Standard Arabic (MSA).

1.2.2.2 Morphological Characteristics

Algerian dialect also has different morphological aspects, consisting essentially of a simplification of some inflections and the inclusion of new clitics as follows:

- The casual endings in nouns and verb moods are lost.
- The indicative mood is used as the default, unlike other moods that are rarely used.
- The dual form and the feminine plural have disappeared and have been absorbed into the masculine plural.
- The first and second person singular forms are conjugated the same way. For example: ``شكرت" can mean both ``I thanked" and ``you thanked."

1.2.3 Grammatical Structures in Algerian Darija

Algerian Darija also shows several grammatical differences compared to Modern Standard Arabic. In this section, we will present some of the main features related to demonstratives, pronouns, verb conjugation, and the structure of verbal sentences.[3]

1.2.3.1 Personal Pronouns in Algerian Darija

Personal pronouns are an essential part of daily communication. In Algerian Darija, they differ from Modern Standard Arabic in both singular and plural forms. The distinction between masculine and feminine is often dropped in everyday speech. The table 1.3 shows these pronouns in Algerian Darija compared to Modern Standard Arabic.[2]

	Singular		Plural
	Feminine	Masculine	Masculine & Feminine
Person 1st	أنا / I	أنا / I	حنا / We
Person 2nd	نتِ / You	نتَ / You	نتوما / You
Person 3rd	هي / She	هو / He	هما / They

Table 1.3: Personal pronouns of Algerian Darija.

Table 1.4 shows how the verb راح ("to go") is used in Algerian Darija. It is the same as **ذهب** in Standard Arabic. In Darija, verbs are easier: there are no endings, and the words are shorter to say. Sometimes, the same word is used for both masculine and feminine, depending on the situation. The difference between singular and plural still exists but is more flexible in daily speech.

Pronoun	**ALG (Algerian Darija)**	**MSA**	**English**
1st Person (Singular)	انا نروح	أذهب	We go
2nd Person (Feminine)	نت تروحي	تذهبين	You go
2nd Person (Masculine)	نت تروح	تذهب	You go
2nd Person (Plural)	نتوما تروحو	تذهبون	You go
3rd Person (Feminine)	هي تروح	تذهب	She goes
3rd Person (Masculine)	هو يروح	يذهب	He goes
3rd Person (Plural)	هما يروحو	يذهبون	They go

Table 1.4: Comparison of the pronouns and the verb 'To go' in Algerian Darija, MSA, and English.

1.3 Demonstratives in the Algerian Darija

Demonstrative pronouns help us show or point to people or things that are close or far, in place or time. In Algerian Darija, these pronouns are different from Modern Standard Arabic in how they sound and how they are built.[2, 3] Often, the same word is used for both masculine and feminine, which makes them easier in daily speech. Table 1.5 shows the most common demonstrative pronouns in Algerian Darija, divided into singular and plural forms.

Singular		Plural
feminine	Masculine	feminine & Masculine
هادي	هادا	هادو
This	This	These
هاديك	هاداك	هادوك
That	That	Those

Table 1.5: Demonstrative pronouns of Algerian Darija.

1.4 Sentence Structure in Algerian Darija

In Algerian Darija, sentences can have different word orders. In Modern Standard Arabic, the usual order is Verb-Subject-Object (VSO). But in Darija, the word order is more flexible. The most common order is Subject-Verb-Object (SVO). Sometimes, the order changes depending

on the situation or style. Table 1.6 gives examples to show these differences using a simple sentence.[2]

Order	Dialect Sentence	English
SVO	لولد راح للمدرسة	The child went to school
VSO	للمدرسة راح لولد	
OVS	للمدرسة لولد راح	
OSV	لولد للمدرسة راح	

Table 1.6: Example of word order in a ALG declarative sentence.

1.5 Interrogative Words in Algerian Darija

Interrogative particles serve essential communicative functions, enabling speakers to elicit information across various domains of discourse. In Algerian Darija, the forms of these interrogatives differ from those in Modern Standard Arabic, both [2]in pronunciation and structure. The table 1.7 below presents the most common interrogative particles and pronouns used in the Algerian dialect, alongside their equivalents in MSA and their meanings in English.

ALG	MSA	English
شكون	من	Who
وين	أين	Where
واشن / واش	ماذا	What
باش	بماذا	With what
وقتاش	متى	When
وعلاش	لماذا	Why
كفاش	كيف	How
شحال	كم	How many

Table 1.7: Interrogative particles and pronouns in ALG and their equivalents in MSA.

Conclusion

This chapter provided a comprehensive overview of Algerian Darija, describing its phonological, morphological, lexical, and syntactic characteristics. We examined the influence of historical language contact with Berber, French, Turkish, and Spanish, as well as regional variation that shapes the dialect's linguistic profile.

Through the analysis of personal pronouns, verb conjugation, demonstratives, interrogatives, and sentence structures, we illustrated the structural simplifications and unique features

of Algerian Darija.

Understanding these linguistic properties is crucial for the development of effective Natural Language Processing (NLP) models and computational tools capable of handling the complexity and richness of Algerian Darija.

Chapter 2: Sentiment Analysis

Introduction

Sentiment analysis is a key technique in understanding emotions, attitudes, and opinions expressed in various communication channels, such as social media platforms, customer reviews, and textual feedback. It provides valuable insights into public sentiment, customer satisfaction, and emotional trends, making it an essential tool in many fields.This chapter explores the fundamentals of sentiment analysis, including its concepts, different levels, methods, challenges, and practical applications across diverse domains.

2.1 Project Context

Sentiment analysis is a broad field with many concepts. In this section, we will mention the levels of sentiment analysis, its challenges, and its techniques:

2.1.1 Levels of Sentiment Analysis

We can perform sentiment analysis on three levels, depending on the task's details, which Provides a range of possibilities for analysis[4].

2.1.1.1 Document-Level Sentiment Analysis

This level of sentiment analysis involves analyzing [5]the sentiment of an entire document, assuming that the review belongs to a single person. The purpose is to identify whether the document's overall sentiment[4] is positive, negative, or neutral.

2.1.1.2 Sentence-level Sentiment Analysis

At this level, each sentence is subjected to sentiment analysis to determine whether it is neutral, positive, or negative. Before that, we have to categorize sentences into objective[5] sentences when there [4]is no sentiment present, for example: "The iPhone 16 Pro has three cameras. Or subjective sentences, where we can use NLP to extract emotion from them, for example: "iPhone 16 is the best phone right now".

2.1.1.3 Aspect Level Sentiment Analysis

Aspect-based sentiment analysis, also known as named entity analysis, focuses on analyzing the sentiment of specific aspects or features of a product or service. For instance, in a product[5] review, sentiment analysis can determine whether the reviewer expresses positive or negative sentiments about aspects such as price,[4] usability, or customer service. For example: "I think that the Fiat 500 is a good car, but when it comes to price, it's overpriced". The sentiment depends on the aspect being considered, as positive, but if we consider the price, it would be classified as negative.

2.2 Applications of Sentiment Analysis

Sentiment Analysis has gained prominence across diverse sectors such as education, health care, business, and government due to its adeptness in deciphering subtle emotional details embedded within textual data. In educational settings, the deployment of Sentiment Analysis facilitates Accurate comprehension of sentiments expressed by students and educators. Through careful examination of feedback channels and identification of technological inclinations, in sights into prevailing sentiment landscapes, particularly evident in educational technology con ferences and MOOCs, are attained[6].

Within the healthcare domain, Sentiment Analysis plays a pivotal role in understanding patient feedback dynamics, providing valuable insights into engagement patterns and early detection prospects within mental health domains [7]. Businesses also leverage Sentiment Analysis's analytical capabilities to discern customer sentiments regarding their offerings. This drives iterative enhancements, boosts customer satisfaction, and strengthens competitive positions [5]. Additionally, in governmental contexts, strategic utilization of Sentiment Analysis

enables the discernment of public sentiment regarding policy frameworks and service provisions. Through the scrutiny of interactions between citizens and government entities on social media platforms such as Twitter, policymakers can extract invaluable insights[8] to inform evidence-based decision-making processes. These diverse applications underscore Sentiment Analysis's indispensable role as a sophisticated analytical tool, entrenched at the forefront of contemporary data-driven inquiry and strategic decision-making paradigms.

2.3 Sentiment analysis challenges

Before discussing the specific challenges of sentiment analysis in the Algerian Arabic dialect, these are some common issues when performing a sentiment analysis generally:

- **Spam:**Spam refers to unsolicited or unwanted messages that may contain irrelevant or misleading information that can affect the accuracy of sentiment analysis. Businesses usually use it to promote product with fake reviews.[9] There is lot of English spam detection research, 5 like wish is a review about massive number of works on this issue. Unlike the Arabic language where is still ongoing research (achieves accuracy of 99%) But for the Algerian dialect, there is only few works like .

- **Review quality:**Not all reviews are helpful, the objective reviews have no emotion and are therefore useless. so only the subjective reviews should be kept. There are also comments[10] that do not express a sentiment about the product, like ' صباح الخير ' (good morning in Arabic) which is classified as positive . A lot of researches globally try to detect subjectivity. In ,they use Bayesian network based algorithm, a later workuse BERT pretrained models for subjectivity detection. Concern Arabic context, in the authors build SAMAR which is a system for subjectivity and sentiment analysis.

- **Sarcasm:**Sarcasm can be a big issue in sentiment analysis since the literal meaning of a sentence may be opposed to its intended meaning. Because it is a type of irony where the[11] speaker means the exact opposite of what he is saying.

  This comment says "The prices are reasonable and affordable to the average citizen" which is sarcasm about expensive prices, and what we notice is the upside-down face Emoji, which many people use for sarcasm, demonstrating the role that emojis can play

in detecting true emotions. Not just the emojis, either. For example, لا ميش غالية خلاص كون زدت شوي ماشي خير The authors of investigate the impact of sarcasm on sentiment analysis and propose hashtag tokenization as a solution to this problem.

- **Negation handling**Because they can change the polarity of a sentence, handling negation word like not, neither, nor, etc. and ماش ما لا,in Arabic Algerian dialect, is important[11] for sentiment analysis. For instance, the sentence كل يوم العيد.

  would be considered a positive sentence, but ماش كل يوم العيد would be considered a negative sentence. Some techniques implicitly exclude negation words since they have a neutral sentiment value in a lexicon so they don't affect the final polarity. However, Negation words can be found in sentences without changing the meaning of the text, which makes it tricky to accomplish this task by reversing the polarity.

- **Domain dependency:**any words can have different meanings depending on the context in which they are used. [12]For example: when referring to the covid test, the word "positive" have a negative meaning. One of the ways to solve this problem is to build domain-specific lexicons like , but this solution is not suitable for Arabic language, because it has a lot of variants and dialects and it is complicated to deal with two generalities at once (dialect and domain), as the authors of said.

- **Data Imbalance:** Imbalanced sentiment analysis datasets skew model accuracy toward majority sentiment classes, limitations minority sentiments.

- **Data Quality and Noise:** Poorly labeled or noisy training data undermines sentiment analysis model effectiveness by introducing inaccuracies and irrelevant information. For example, noisy customer review data may include irrelevant information or misleading sentiments.

- **Multilingualism:** Analyzing sentiment across languages presents challenges due to structural, grammatical, and expressive differences. For instance, translating sentiments accurately between languages can be complex.

2.4 Related Works

Many studies have been conducted in sentiment analysis for the English language, while only a few have focused on Arabic, particularly the Algerian dialect. This section presents a chronological overview of the most relevant works, taking into account the techniques used.

- One of the first works on sentiment analysis of the Algerian dialect is on 2016. The authors use lexicon-based approach, with lexicon of 3k different words adapted to Algerian dialect from Egyptian work,[13] then they test it on 7k comments, and get accuracy of 53.3% with basic preprocessing then improve it to 79.13% with Arabization and translation of words from other alphabets and languages, stemming with khoja Stemmer 3, and handling non-included words with polarity computation module.

- Later on 2018, the authors of,[14] create lexicon of 4800 words translated from another English lexicon, that lexicon help them to automatically annotate corpus of 8k messages (4k Arabic 4k arabizi). Then, they use it to train with machine learning algorithms (SVM, NB, LR, DT, RF) with the word features (BOW and doc2vec). As a result, they get 68% as accuracy on Arabic and 66% on arabizi after performing tests on external dataset (results better in internal ones).

- After that, on 2019, the authors of [2] use machine learning (SVM with TF-IDF) and deep learning (LSTM with word2vec trained on 1.4 gigabytes data), both trained on 50k Arabic comments (no arabizi), the SVM get 86% accuracy while the deep learning model get 82%.[15] In 2020, the authors of , build TWIFIL, a corpus of 9k Algerian tweets, and train it with SVC (adaptation from SVM) and the frequency as data representation, which made them reach an accuracy of 69.53%. The deep learning techniques overperform the machine learning ones.[15] CNN, LSTM and BERT get 76%, 75% and 68% respectively

  2020, the authors of , build TWIFIL, a corpus of 9k Algerian tweets, and train it with SVC (adaptation from SVM) and the frequency as data representation, which made them reach an accuracy of 69.53%. The deep learning techniques overperform the machine learning ones.[16] CNN, LSTM and BERT get 76%, 75% and 68% respectively.

  on 2022 in [17], the authors delete the stop words with NTLK package, and tokenize the words with Camel tools [36], and they use TF-IDF as feature selection with machine

learning algorithms (SVM, BNB and MNB). As a result, they achieve an accuracy of 67%, 66% and 64 respectively. With the deep learning they prefer word2vec instead of TF-IDF, after using LSTM and MarBERT (a Moroccan pre-trained model), they achieve 79% and 82% on accuracy.

- In the same year, the authors of used FastText as a feature extraction technique (a new extension of Word2Vec). [18]After they found it more suitable to the Algerian dialect, because they need the closest word similarity and not the similarity across the context, they collected 50k comments, manually annotated them, and removed manually constructed stop words from them. Then, they train LSTM and CNN. The results show that LSTM achieves 82% and LSTM+CNN achieves 79% in accuracy.

- 2021 was a rich year in the field of NLP for the Algerian dialect, where the authors of announce the first pre-trained transformers module specialized in the Algerian dialect, the authors of this work fine-tuned their model to the sentiment analysis task, and test it on Twifil and Narabizi datasets, it gives good results in comparing to AraBERT and MarBERT . On Twifil The obtained results are: AraBERT: 73.8%, MarBERT: 80.6% and DziriBERT: 80.5%. On Narabizi: AraBERT: 73.8, MarBERT: 80.6 and DziriBERT: 80.5.generally in the other metrics (Precision, Recall, F1-score).[19]

- On a paper published in 2023 , the authors collect and label 16k tweets (Arabic + arabizi) about the Algerian social movement. They train them with advanced machine learning techniques (Linear SVC, Multi-NB, Multi-LR and DT with TF-IDF or BOW as features extraction), and they achieve accuracy of 67%, 62%, 68% and 58% respectively. With the pre-trained models (DziriBERT, AraBERT and XLM-R) they get 67%, 66% and 63% respectively.[20] We see that DziriBERT achieves the best accuracy results comparing the other pre-trained modules, and it also achieved the best results generally in the other metrics (Precision, Recall, F1-score).

gray!30 **Ref**	**Techniques**	**Feature extraction**	**Datasets**	**Results (accuracy)**
[17] 2022	SVM, BNB, MNB, LSTM, BERT (MarBERT)	Word2vec, NLTK stop words, Camel tokenization	20k tweets (Pos, Neg, Neu)	SVM: 67.5% MNB: 66.65% BNB: 64.5% LSTM: 79.04% BERT: 82.36%
[15]2019	SVM + TF-IDF, LSTM + CBOW	Word2vec (1.4 GB corpus)	49k items (10k collected) (Pos, Neg)	SVM: 86% LSTM: 81%
[13] 2016	Lexicon-based	3k words + Arabization + translation + stemming	7k comments (Pos, Neg)	Basic: 53.3% Advanced: 79.13%
[20] 2023	LSVC, MNB, MLR, DT, RoBERTa, AraBERT, DziriBERT	BoW and TF-IDF	16k tweets (Pos, Neg, Neu) about Algerian social movement	LSVC: 67% MNB: 62% MLR: 68% DT: 58% RoBERTa: 63% AraBERT: 66% DziriBERT: 67%
[16] 2020	SVC, MLP, CNN, LSTM, BERT	Frequency (better than BoW, TF-IDF)	TWIFIL 9k tweets (Pos, Neg, Neu)	SVC: 69.5% MLP: 73.4% CNN: 76% LSTM: 74% BERT: 68%
[19] 2021	DziriBERT, AraBERT, MARBERT	WordPiece Tokenizer	1.2M tweets (Pos, Neg, Neu) fine-tuned on TWIFIL + Narabizi	TWIFIL: AraBERT: 73.8% MarBERT: 80.6% DziriBERT: 80.5% NArabizi: same
[18] 2022	LSTM, CNN	FastText	50k tweets (Pos, Neg) Manual annotation	LSTM: 82% LSTM+CNN: 79%
[14] 2018	SVM, NB, LR, DT, RF	BoW (Arabic), Doc2Vec (Arabizi)	8k messages: 4k Arabic, 4k Arabizi (Pos, Neg)	Arabic: SVM: 63%, NB: 67%, LR: 68%, DT: 59%, RF: 61% Arabizi: SVM: 63%, NB: 66%, LR: 66%, DT: 55%, RF: 56%

Table 2.1: Related Work

2.5 Synthesis and discussion

From the previous works, we deduce that we can perform sentiment analysis in different ways and techniques, and that this field is achieving continuous rapid development in order to improve results and overcome challenges. For the Algerian dialect, some of the works focus only on sentences written in Arabic letters [17] and[15] and [18], the others trait also the sentences and words written in Latin letters on the pre-processing step in a separated dataset with the translation method [29] and [14], where first they detect English and French terms and then translate them into Arabic language, then they automatically consider the rest of Latin words as arabizi and transliterate them to the Arabic language, with hard-coded functions or with the help of some ready packages. But in the recent works [17], [20] and [20], the researchers use multilingual pretrained model that can trait multiple languages and alphabets at the same time. To transfer the words to a set of features that the algorithms can use, some of the works use simple methods like the TF-IDF technique [20] and [16], but most of them use Word2Vec after training it on a large amount of data[17] and [15]. Recent works use fast text, which is an extension of Word2Vec, and achieve good results [18]. We also note that the size of the dataset is critical for achieving decent accuracy. In[13],[20], [35] and [24], where the number of comments is less than 16k, the accuracy is at most 79%. However, in [17], [15] and[18], where the number of comments exceeds 20k, the accuracy is always greater than 80%. In machine learning, we noticed that SVM is the most commonly used technique, surpassing Naïve Bayes in [17] and [20]. Machine learning has other techniques like logistic regression and decision tree, and it is hard to tell which one gives better results. Deep learning techniques generally give better results than machine learning in the case of using the same feature extraction technique in [17] and [16], and we observe that LSTM give better results than RNN or CNN [16]and[18]. The pre-trained models generally achieve better results than other models [17]and [20], in the Algerian dialect context, DziriBERT always gets higher results than other models due to the fact that it specialized in the Algerian dialect [20] and [19], which makes it the current state of the art of sentiment analysis of the Algerian dialect.

Conclusion

This chapter provided an overview of sentiment analysis, covering its definition, levels, and real-world applications. We discussed common challenges such as sarcasm, negation, domain dependency, and multilingual data, with a focus on the Algerian dialect. A review of recent related works highlighted the evolution of techniques from lexicon-based methods to deep learning and specialized pretrained models like DziriBERT, which currently leads in accuracy for Algerian sentiment analysis. Overall, this chapter sets the foundation for understanding sentiment analysis and its application to Algerian dialect data.

Chapter 3: Theoretical Background

Introduction

After introducing concepts and related works in the previous chapter, this chapter will focus on our contributions to the subject of sentiment analysis of the Algerian dialect. We will begin by discussing the theoretical proposal of our solution. Subsequently, we will proceed with the implementation of our solution, providing a detailed explanation of the involved steps.

3.0.1 Deep learning:

Deep learning is a specialized area within artificial intelligence (AI) that relies on advanced neural networks with multiple layers to analyze complex datasets. Instead of following fixed rules, this technique enables systems to recognize patterns and make decisions autonomously by learning from the data itself. It has proven to be particularly effective in areas like object detection, voice recognition, and machine translation. Deep learning is being adopted across many sectors, including healthcare for enhancing diagnostic accuracy, retail for crafting personalized experiences for customers, and the automotive industry to improve vehicle safety and operational efficiency[21] An illustration of a deep neural network's architecture is shown in figure below Some of the key techniques used in deep learning include.

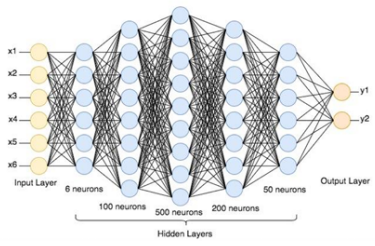

Figure 3.1: Deep neural network architecture

1. **Long Short-Term Memory Network (LSTMN):** Long Short-Term Memory Network is type of recurrent neural network (RNN) designed for modeling sequential data. It handle dependencies and retain information over time using memory cells. LSTM capture long-term relationship by remembering or forgetting at each step. It is used in speech recognition, machine translation, sentiment analysis, and more[22].

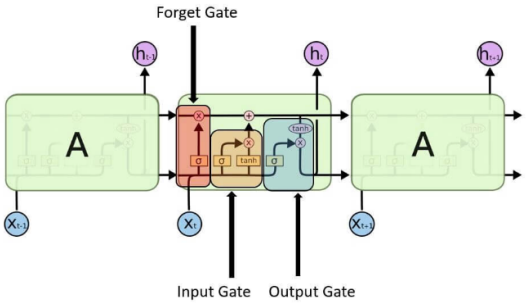

Figure 3.2: Long Short-Term Memory Network Architecture

2. **Bidirectional Long Short-Term Memory (BiLSTM)** is an advanced Recurrent Neural Network (RNN) architecture used in Natural Language Processing (NLP). It processes sequence data in both forward and backward directions, allowing it to capture context from both preceding and succeeding words, thus providing a richer understanding of the sequential input[23].

3. **Recurrent Neural Networks (RNNs):** Recurrent Neural Networks (RNNs) are connectionist models designed for processing sequential data. They are feedforward neural networks with recurrent edges that span adjacent time steps, allowing them to retain a state representing previous steps in the sequence. This structure allows RNNs to model dependencies across sequence steps and time dependencies on multiple scales. RNNs are widely used in tasks involving sequences, such as time series prediction, video analysis, speech processing, and text generation[24].

4. **Convolutional Neural Networks (CNNs):** Convolutional Neural Networks are used to analyze visual data like images and videos. They are great at automatically learning and extracting features from the input data. CNNs excel at capturing complex patterns and hierarchical representations within the data. This makes them highly effective for tasks such as image classification, object detection, and images segmentation[25]. The key complements of a CNN include convolutional layers, pooling layers, and fully connected

layers. Here's brief explanation for each:

- **Convolutional layers:** these layers have learnable filters that slide over input data to detect features like edges, textures, or shapes. They produce feature maps
- **Pooling layers:** Used to downsample feature maps, reducing dimensions. Max polling preserves prominent features.
- **Fully Connected layers:** transform high-level features into desired output format. An example of a convolutional neural network including multiple convolution and pooling layers is shown in3.3

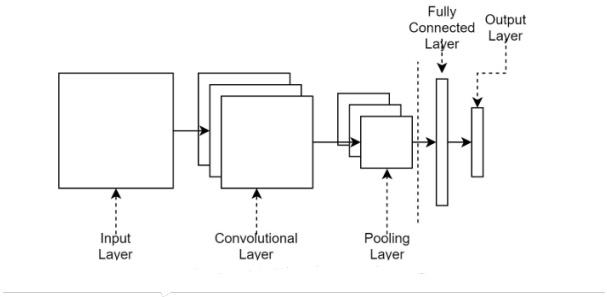

Figure 3.3: Convolutional Neural Network (CNN) Architecture.

5. **Transformers and Pre-trained Language Models:**

Transformers are deep learning techniques that revolutionize the natural language process ing (NLP) field. They have first been introduced in the paper (Attention is All You Need) [26]. Since then, they have attracted much interest due to their ability to capture long-range dependencies in text data[26]. Transformers use a self-attention mechanism that enables them to process words in parallel, focus on the essential pieces of information, and filter the less important ones, making them highly effective at processing enormous amounts of text data. Pre-trained Transformer based language models, such as BERT [27], GPT-4 [28], and RoBERTa [29], have become increasingly popular in sentiment analysis classification [30]. These models are pre-trained on large datasets and can be fine-tuned on specific tasks, such as sentiment analysis. They have achieved state-of-the-art results on NLP benchmarks and require less data and computational resources than traditional deep learning techniques. The figure 3.4shows the original transformers architecture

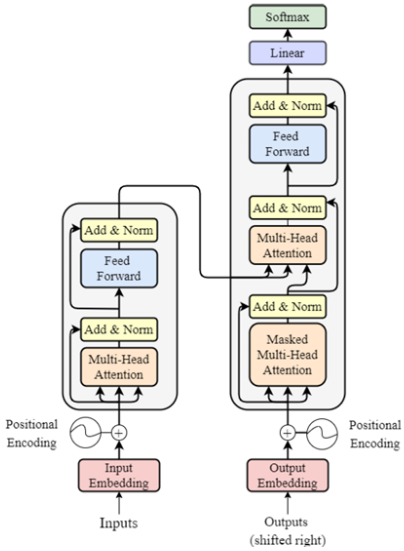

Figure 1.4: **Original transformers architecture**

Figure 3.4: Original transformers architecture

Furthermore, we will introduce some of the more suitable pre-trained modules for sen timent analysis of the Algerian dialect.

- **(Bidirectional Encoder Representations from Transformers)** is a pre-trained language model that uses a transformer architecture to learn the context of words in a sentence. The model is trained bidirectionally on a huge corpus of text data using a Masked Language Modeling task (MLM), which masks 15% of the words in the input and tries to predict them from the remaining words. BERT can also be fine-tuned for a variety of downstream applications, such as sentiment analysis [31].

- **ARABERT** is a pre-trained language model that was created exclusively for Arabic language processing. The model is built on the BERT architecture and was trained on a huge corpus of Arabic text data (70 million sentences) using the BERT base configuration (12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximumsequencelength, and a total of 110M parameters) [32]. ARABERT fine-tuned on various three tasks: Name Entity Recognition, Question Answering and Sequence Classification that we need to do sentiment analysis.

- **DZIRIBERT** is a pre-trained language model that is specifically designed for the Alge rian Arabic dialect. The model is trained on 1.2 million Algerian tweets and uses the same architecture of BERT base, and can be fine-tuned on various downstream tasks, including sentiment analysis [33].

3.1 Machine learning

Machine learning technology is crucial in modern society for tasks like web searches, content filtering, and e-commerce recommendations. It is also used in consumer products like cameras and smartphones for tasks like image identification, speech transcription, and search results selection. Deep learning techniques have surpassed other methods in various domains. Supervised learning is the most common form, while trainable multilayer networks were initially intended for pattern recognition. Convolutional neural networks (ConvNets) and recurrent neural networks (RNNs) are popular for tasks involving sequential inputs[34].

3.1.1 How does machine learning work?

Machine learning operates by starting with data. A machine learning model will be trained using the data, which has been gathered and prepared to serve as training data. The performance of the application improves as the data volume increases [35]. Once the data is prepared, programmers select a machine learning model and provide the data for training. The computer model then learns from the data, identifying patterns and making predictions. When presented with new data, the trained model utilizes its learned knowledge to generate outputs or predict outcomes. The quantity and quality of training data impacts the accuracy of these predictions. A system can build more complex models and predict more accurately when it has access to high-quality data sets.

- **Supervised learning:** is the most common form. It uses labeled data sets, meaning each input is paired with its corresponding correct output or "target". The machine learns the relationship between the inputs and these desired targets[36].

- **Unsupervised learning:** involves the program looking for patterns in unlabeled data, where there are no predefined targets. The goal is to find inherent structures or groupings within the data.

- **Reinforcement learning:** trains machines through trial and error by using a reward system to guide the model towards taking the best actions. After the model has been trained on the data, its performance is evaluated on a separate test set of examples

it has not seen before. This step measures the model's generalization ability—its capacity to produce correct or sensible outputs on new, unseen inputs[37]. Ultimately, machine learning systems can perform various functions, including being descriptive (explaining past events), predictive (forecasting future outcomes), or prescriptive (suggesting actions to take) based on the data and patterns they have learned.

3.1.2 Machine learning applications:

Each day, machine learning advances quite quickly. Without even realizing it, we utilize machine learning in our daily lives through applications like Google Maps, Google Assistant, Alexa, etc. Here are some of the trendiest applications of machine learning: Here are some of the trendiest applications of machine learning[38]:

- Image Recognition, Speech Recognition
- natural language processing (NLP)
- Product recommendations
- language translation
- financial services
- autonomous vehicles
- Game playing

3.2 Machine learning algorithms

The many kinds of machine learning use a number of algorithms. These are a few algorithms:

3.2.1 Support Vector machine (SVM)

Support Vector Machines (SVMs) are among the most prominent and effective machine learning models used today. They operate under a supervised learning framework and

are designed to perform both classification and regression tasks. Beyond simple linear classification, SVMs are capable of addressing complex, non-linear problems by applying the kernel trick, a method that allows data to be transformed into a higher-dimensional feature space where it becomes more separable. The fundamental goal of SVMs is to find the optimal hyperplane that separates different classes with the widest possible margin, ensuring that the boundary is as far as possible from the nearest data points of each class, which in turn helps to minimize classification errors.[39]

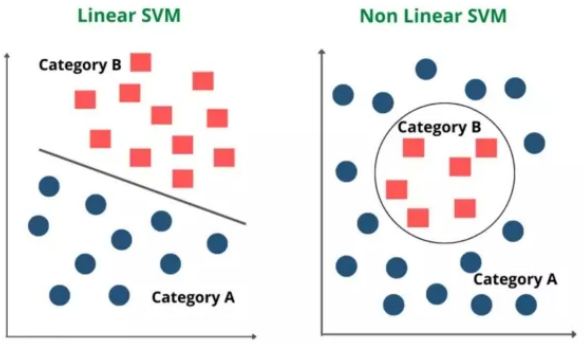

Figure 3.5: Types of SVM separation.

3.2.2 The Naïve Bayes

The Naïve Bayes classifier is a supervised machine learning algorithm. It is a classification technique that relies on Bayes' Theorem and assumes independence among predictors. In simpler terms, it assumes that the presence of one feature in a class is not related to the presence of any other feature. Naive Bayes is commonly used in the text classification industry. It is primarily employed for tasks such as clustering and classification, where the main focus is on determining the conditional probability of an event occurring based on the available features [39]

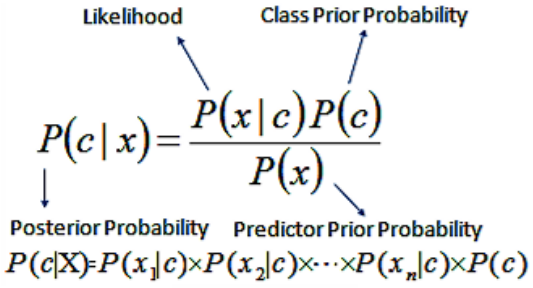

Figure 3.6: Navie Bayes .

3.2.3 Decision Tree (DT)

Decision tree is a type of supervised learning algorithm used in machine learning and decision- making processes. It is a tree-shaped diagram that represents a series of decisions and their possible consequences. It follows a hierarchical structure that includes a root node, branches, internal nodes, and leaf nodes. Decision trees can be used for both classification and regression problems. In sentiment analysis, decision trees can be used to classify text data based on their sentiment polarity by recursively splitting the data based on the most informative features[40].

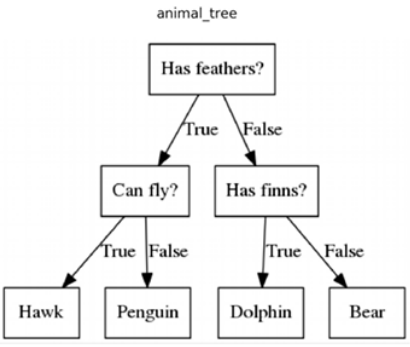

Figure 3.7: Decision Tree example .

3.2.4 Random Forest (RF)

Random Forest is a machine learning algorithm that was first proposed by Leo Breiman and Adele Cutler in 2001 [41]. It is an extension of the decision tree algorithm and is widely used for classification and regression tasks. The algorithm works by creating multiple decision trees and combining their predictions to produce a final output.

3.2.5 Artificial Neural Network (ANN)

Neural Network learning algorithms are computational systems that process input data into desired outputs.[42] Inspired by biological neurons in the human nervous system, neural networks were developed by Alexander Bain and William James. Each neuron has an activation function and weights that determine its connection to other neurons. Layers of neurons process incoming input, with hidden layers performing intermediate processing. The output layer receives the output of the last hidden layer, creating the final output. Neural networks are popular tools for model recognition, classification,

and prediction, and can be developed using supervised, unsupervised, and reinforcement learning approaches. They have been successful in various domains, including image and speech recognition, natural language processing, and predictive analytics[43].

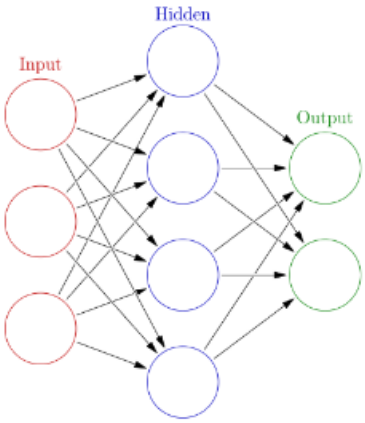

Figure 3.8: Artificial neural network layers.

3.3 Datasets Overview

In this section, we present an overview of the datasets used for sentiment analysis in the Algerian dialect, including both existing datasets from previous research and our custom-built dataset. Each dataset has been collected manually from specific sources and covers different topics such as telecommunications, electronics, and general discussions. These datasets are essential for training and evaluating machine learning and deep learning models in the context of Algerian dialect sentiment analysis

3.3.1 Djezzy Dataset by Salhi et al. (2021)

In their study, Salhi et al. (2021)[44] focused on analyzing tweets related to the Algerian mobile operator Djezzy to assess customer satisfaction and the company's e-reputation. The dataset comprised tweets in Arabic, French, and English, collected from Twitter. The authors performed pre-processing steps, including filtering out irrelevant words and tokenization, to retain essential information for accurate sentiment analysis. Subsequently, they applied machine learning algorithms such as Support Vector Machine (SVM) and Logistic Regression for supervised classification of sentiments. A notable aspect of their approach was the deployment of these algorithms on a cloud infrastructure, enhancing

execution time and accessibility. The solution supported multilingual data processing, accommodating Arabic, English, and French tweets.

3.3.2 Brandt DZ Dataset by Chader et al. (2019)

The Brandt DZ dataset was introduced by Chader et al. (2019) to address sentiment analysis challenges in the Algerian dialect, especially written in Arabizi (Algerian Arabic using Latin script). This dataset contains 6,738 comments manually annotated into positive and negative sentiment classes. The comments were collected primarily from social media platforms, reflecting authentic user opinions expressed informally in the Algerian dialect. The authors highlighted the difficulties posed by the non-standard spelling and variability of Arabizi, making this dataset particularly valuable for developing robust sentiment classification models for Algerian dialect texts[45].

3.3.3 TWIFIL Dataset by Mataoui et al. (2020)

Mataoui et al. (2020) created the TWIFIL dataset, which includes 6,001 manually annotated comments covering general topics. This dataset is primarily focused on informal social media text, making it relevant for sentiment analysis tasks in real-world scenarios. The annotation process involved categorizing comments as positive, negative, or neutral sentiments. The dataset supports the development of models that handle the complexity of dialectal Arabic variations and the informal nature of social media language[46].

3.3.4 Mataoui Dataset by Mataoui et al. (2020)

Another dataset provided by Mataoui et al. (2020) contains 11,303 comments annotated manually and covering a variety of general topics. This dataset further expands the resources available for dialectal Arabic sentiment analysis and offers a rich source of diverse sentiment expressions from social media. Its large size and manual annotation ensure high quality for training deep learning models[47].

3.3.5 Our Custom Dataset

Our dataset consists of 27,774 manually annotated comments related to the automotive industry in Algeria. The data was collected from various social media platforms including

Facebook, Twitter, and YouTube. We focused on gathering comments expressing opinions about cars, which allowed us to build a large, domain-specific dataset to improve sentiment analysis performance in this sector. The annotation process was carefully designed to ensure the accuracy and consistency of sentiment labels[48].

Conclusion

In this chapter, we presented the theoretical foundations necessary for understanding sentiment analysis approaches, particularly those based on deep learning. We provided a detailed overview of key architectures such as LSTM, BiLSTM, RNN, and CNN, as well as transformer-based models like BERT and DziriBERT, which have revolutionized the field of natural language processing. We also highlighted the importance of contextual word representations and the impact of pre-trained models on improving classification performance, especially in dialectal contexts such as the Algerian dialect. These theoretical insights form the basis for the proposed methodology, which will be discussed in the following chapter.

Chapter 4: Methodology

Introduction

In this chapter, we present the detailed methodology adopted for preparing the Algerian dialect sentiment analysis dataset and developing appropriate models. We began by collecting data from diverse sources to ensure comprehensive linguistic representation. This was followed by extensive preprocessing steps to handle the specific challenges of Algerian dialect and Arabic text processing. Various feature extraction techniques such as TF-IDF, Word2Vec, and FastText were applied to convert textual data into numerical representations suitable for model training. We then explored and trained multiple models categorized into three main groups: traditional machine learning algorithms, deep learning models, and pre-trained transformer-based models. Finally, the adopted evaluation metrics and performance assessments were discussed to comprehensively analyze the models' behavior.

4.1 Theoretical proposal

For our theoretical proposal, we will begin by reviewing the existing datasets for the Algerian dialect and comparing them with our dataset. Next, we will provide a detailed description of how we collected and preprocessed our dataset. We will then introduce various approaches to feature extraction from the text data, such as TF-IDF vectorizer, Word2Vec, and FastText. Finally, we will present the machine learning and deep learning techniques that we employ for analyzing sentiments in the Algerian dialect. To achieve our objective, we followed the following process.

5.1 **Data Collection and Annotation:** We collect and annotate our sentiment analysis dataset, and create a large corpus for training word embedding models such as Word2Vec and FastText.

5.2 **Data Pre-processing:** Once we have collected enough data, we proceed with cleaning and pre-processing the text. Our goal is to eliminate noise and address Algerian dialect variations as much as possible to ensure better understandability. A detailed explanation of this process is provided in the corresponding section.

5.3 **Feature Extraction:** We convert the pre-processed text data into numerical features that can be used to train our model. This involves employing well-established feature extraction techniques such as TF-IDF (Term Frequency-Inverse Document Frequency), as well as word embedding methods like Word2Vec and FastText.

5.4 **Model Training:** We select suitable techniques for the sentiment analysis task, including machine learning algorithms, deep learning architectures, and pre-trained transformer models. We then proceed to train the model on our data.

5.5 **Model Evaluation:** After properly evaluating the trained model using metrics such as accuracy, F1 score, precision, and recall, we proceed directly to the model deployment phase.

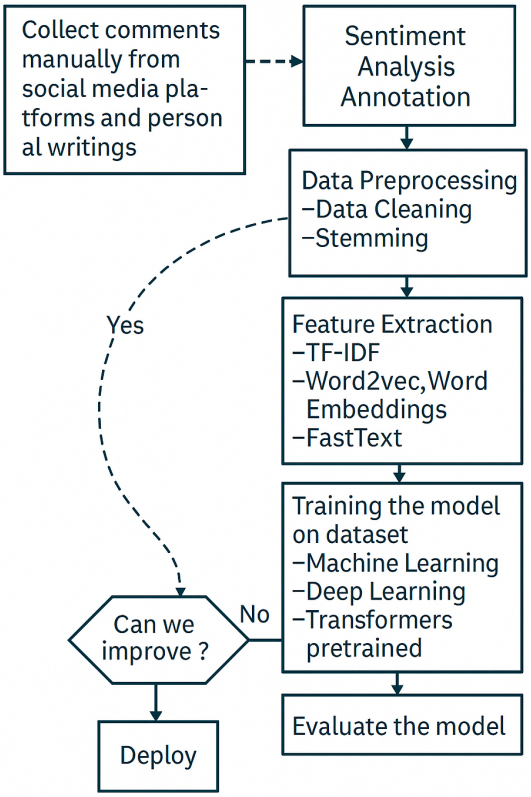

Figure 4.1: Diagram illustrates the steps of our work

.

4.2 Data

In our work, we need two types of datasets. The first is the sentiment dataset, which consists of sentences classified into different sentiment categories such as Negative, Positive, and Neutral. The second dataset is a large collection of texts that will be used for word2vec and FastText models to generate word representations for both Arabic and Algerian words.

4.3 Data Collection

We decided to shift our focus toward **topic-based sentiment analysis** instead of general sentiment analysis, aiming for more precise and context-aware results.

For this project, we **manually collected data** in Algerian Arabic dialect from **social me-**

dia platforms such as Facebook and Twitter, in addition to **personal written content**. The dataset includes a mix of Algerian dialect, Modern Standard Arabic, and even French words, reflecting the linguistic diversity of the Algerian online community.

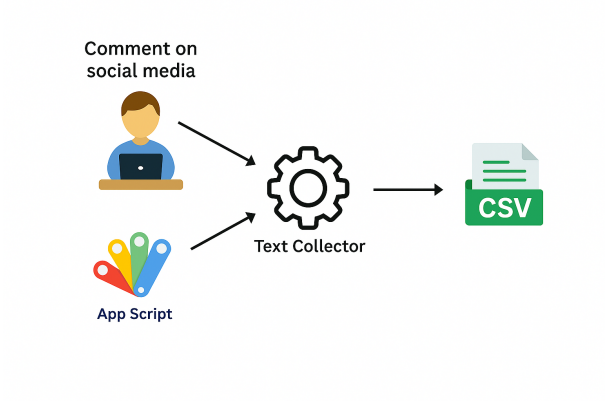

Figure 4.2: Dataset collection .

.

We successfully **collected over 40,000 comments**, which were organized into a structured file to be used in the subsequent stages of preprocessing and sentiment classification.

6. **Sentiment Analysis Annotation:** Out of the 40,000 collected comments, we manually classified 15,774 comments into positive, negative, or neutral categories using an Excel sheet. During this classification process, we also identified and removed spam comments while observing certain characteristics and challenges within our dataset. These observations and insights were taken into consideration during the pre-processing phase of our analysis.

7. **Data analysis:** To do good sentiment analysis, we need first to deeply understand our data. So, we did some statistics that help us to identify the possible challenges we will face "In Figure 4.7, we present the distribution of classes in our dataset, showcasing the percentage and the count for each class. We observe that the data is unbalanced, with the negative sentiment class being the most dominant, containing 9,599 comments (53.8%). Following that, the positive class has 7,165 comments (40.2%), while the neutral sentiment class has 1,065 comments (6.0%)."

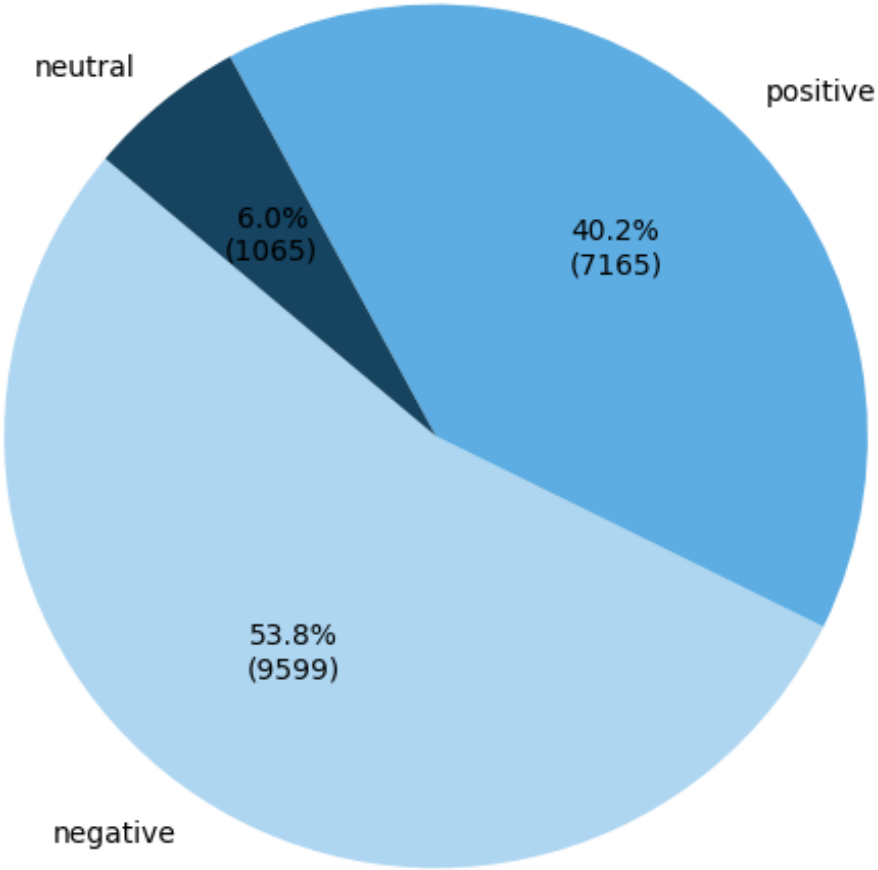

Figure 4.3: Pie chart of the percentage of every sentiment label(positive,negative, neutral)

"In our work, we need a balanced dataset to make a model that does not prefer one label over the others. So, we removed the shortest comments from the classes that have the majority (the negative and neutral comments in our case). Thus, the size of the dataset decreased to 15,110 comments. Figure 4.4shows the equal percentages between the three classes in our dataset, with negative comments at 34.9% (5290 comments), positive comments at 37.6% (5702 comments), and neutral comments at 27.4% (4158 comments)."

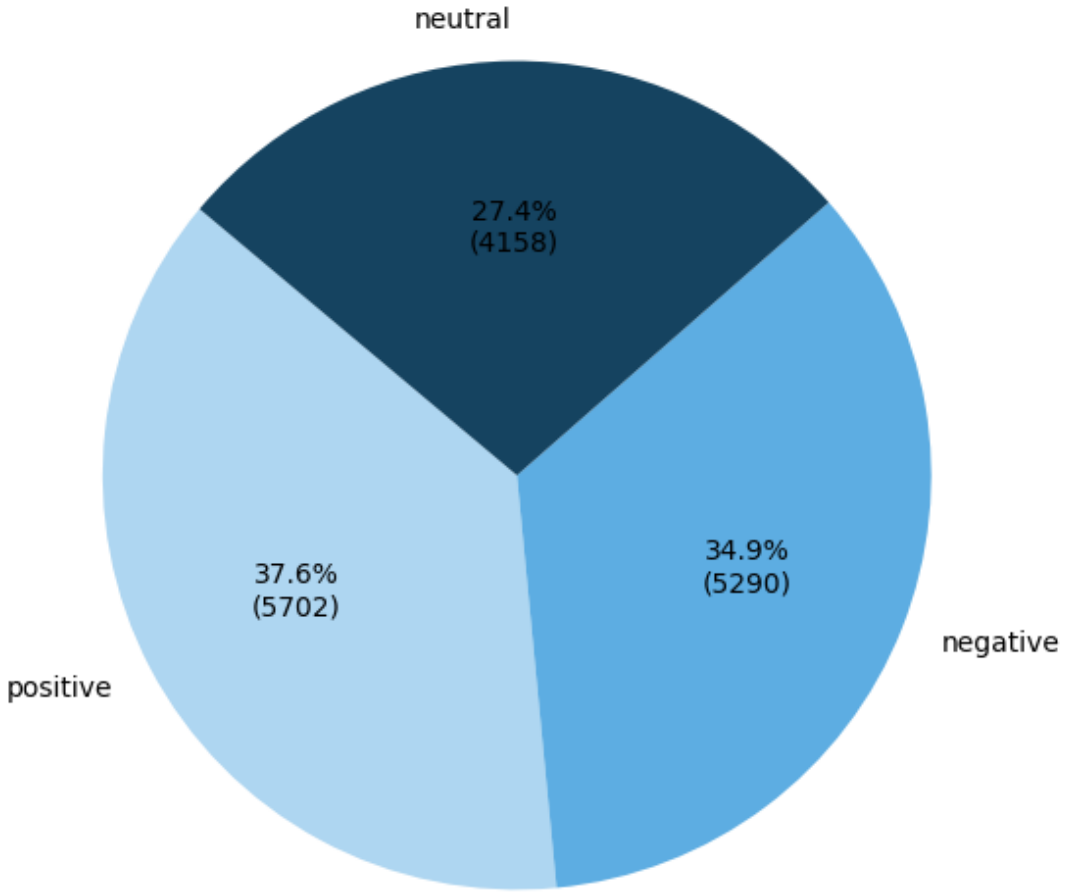

Figure 4.4: Pie chart of the percentage of every sentiment label after making them equal

To know the most common words in every class, we use the n-grams technique. Table 4.1 shows the most common words in every class and the entire dataset, while Table 4.2 shows the most common combinations of two words.

Positive Word	Pos. Freq.	Negative Word	Neg. Freq.	Neutral Word	Neu. Freq.	Entire Dataset Word	Entire Dataset Freq.
شكراً	850	لا	1200	لكن	150	هذا	2500
جيد	780	سيء	1150	ربما	140	من	2300
ممتاز	720	مشكلة	1080	ممكن	130	لا	2200
الله بارك	650	حزين	990	أعتقد	120	على	2100
أحسنت	590	صعب	920	صحيح	110	هو	2000

Table 4.1: Most common unigrams by class.

Positive Bigram	Pos. Freq.	Negative Bigram	Neg. Freq.	Neutral Bigram	Neu. Freq.	Entire Dataset Bigram	Entire Dataset Freq.
لك شكراً	320	أحب لا	550	الممكن من	70	هو هذا	800
جداً جيد	280	جداً سيء	500	أن أعتقد	65	سبيل على	750
رائع عمل	250	جيداً ليس	480	رأيي في	60	المهم من	700
الله بارك	210	بالحزن أشعر	420	جهة من	55	هذه في	650
التوفيق كل	190	كبيرة مشكلة	380	متأكد غير	50	يوجد لا	600

Table 4.2: Most common bigrams by class.

4.3.0.1 Word embeddings corpus:

For the word embeddings corpus, we collected 274MB of text to train Word2Vec and FastText model, which involved merging six different datasets into a single CSV file. The first dataset is 148K comments that we collected from Algerian YouTube channels with YouTube API and App Script, and it contains our sentiment analysis dataset. The second dataset is the Arabic Company Reviews group (Arabic)1. The third dataset is Arabic 100k Reviews, a freely available Arabic collection comprising reviews of hotels, books, movies, products, and airlines from various Arabic news sites2. The fourth dataset is named Arabic reviews 3, an open-source Arabic dataset collected from multiple sources, consisting of reviews about BookMedia hotels and airlines. The fifth consists of 5,870 free Arabic news articles from AlJazeera website 4. Finally, the last dataset is the Arabic Goodreads Reviews, 2013 .

4.3.1 Preprocessing

The preprocessing phase plays a crucial role in enhancing the quality of input data for sentiment analysis models. This leads to improved accuracy in the generated results and faster training of the models. We put our approach of preprocessing to overcome the challenges of sentiment analysis in the Algerian dialect and to generate suitable data for every technique. This section will describe the three levels of our preprocessing approach.

4.3.1.1 Data Cleaning

It involves removing irrelevant or noisy elements from the text. Specific data-cleaning steps were undertaken to ensure the cleanliness of the dataset. These steps include normalizing the Unicode, removing HTML tags (like
), punctuation marks (like `,' `<,' `>), and non-alphabetic characters (like currency symbols, @, and #) that could interfere with sentiment analysis algorithms. Unlike other punctuation marks, we decide to keep the question mark (¿), because it is used a lot in sentences that express a neutral sentiment (Table 2.2 et 2.3). We decided to keep the emojis after separating them with white space. Because when the model treats them as individual tokens, they help to detect the true meaning of the text that is occasionally expressed by emojis. The previous steps are common in the preprocessing of the sentiment analysis of any language. However, in the Arabic and Algerian cases, we need to do extra steps in data cleaning:

- Remove Arabic diacritics (Tashkeel ((تشكيل)
- Keep only one letter in the case of two consecutive letters because many users repeat the letters as an exaggeration, like writing 'هايل' as 'هاااال'.
- Trait unique Arabic letters and unify the way of writing them, which is different from one person to another, transform 'ا آ أ إ' to 'ا', 'ى' to 'ي','ؤ','ئ', 'ي' to 'ئ', 'ة' to 'ء' and 'ة' to 'ه'.
- Remove the stop words, which are common in sentences but do not add any negative or positive meaning to the text. Removing them reduces the training duration without affecting the accuracy of the model.

4.3.2 Word Cloud

The word cloud is a visual representation of the most frequent words in a set of text data. The most frequent words generally appear in larger font and in the foreground, while less frequent words appear in smaller print and in the background.

Figure 4.5: Word cloud of top 200 words

4.3.3 N-grams

refers to contiguous sequences of n words extracted from text for language processing and analysis. An n-gram can be as short as a single word (unigram) or as long as multiple words (bigram, trigram, etc.). These n-grams capture the contextual information and relationships between words in a given text.

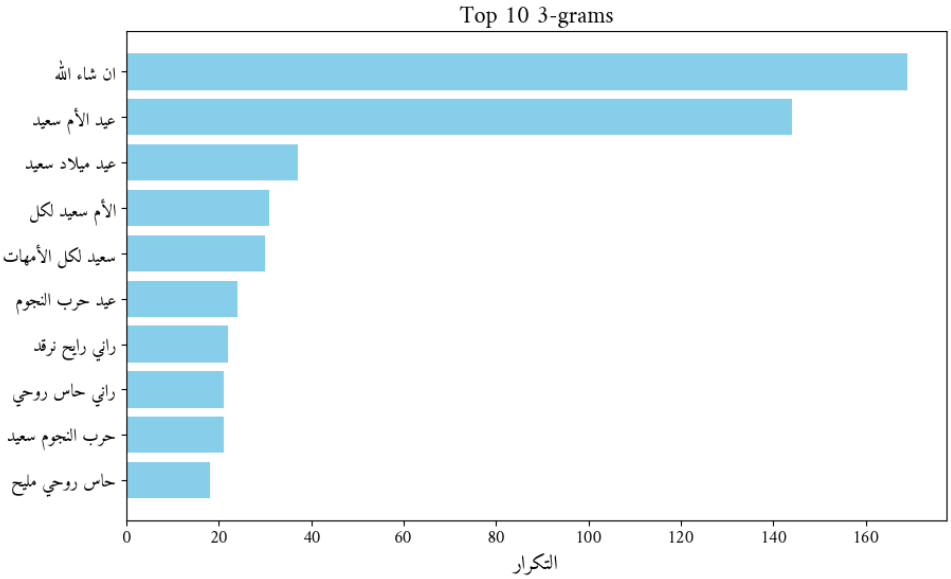

Figure 4.6: Top 10 3-grams

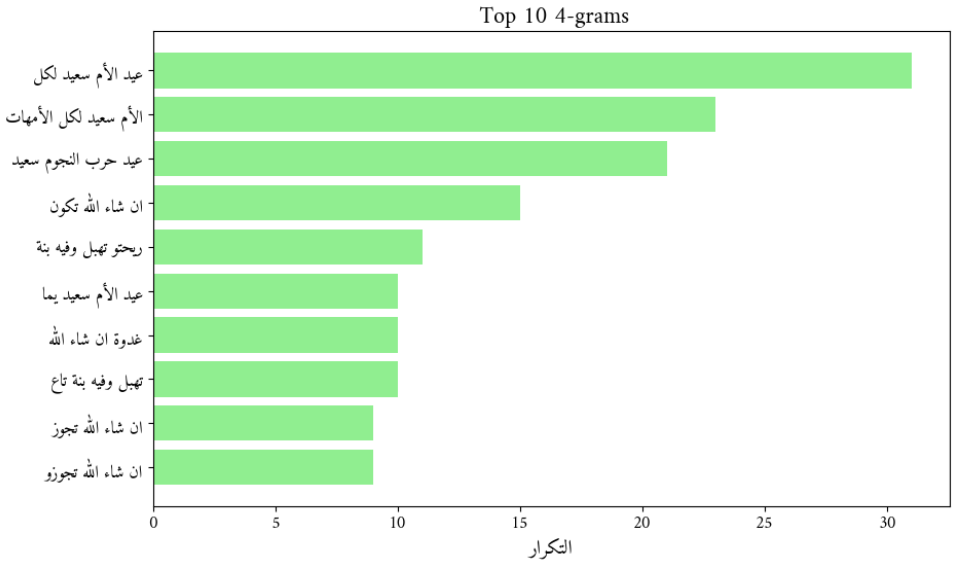

Figure 4.7: Top 10 4-grams

4.3.3.1 Stemming

The stemming procedure aims to change a word into its root form. The primary goal of the stemming procedure is to eliminate all potential affixes, which reduces the word's complexity and the number of characteristics and tokens in corpora. In our approach, we first try to detect if the word is in the Modern Standard Arabic (MSA) or the Algerian dialect. Then, we perform the stemming only on the MSA with one of the available stemmers and leave the Algerian words as they are due to the lack of specialized stemmers for the Algerian dialect.

4.3.4 Feature extraction

In order to extract numerical features from the text for training the model, we employ multiple techniques and compare their performance to determine the most suitable approach for sentiment analysis of the Algerian dialect. In this section, we will describe our approach for feature extraction and explore different techniques used: TF-IDF, Word2Vec, FastText, and BERT.

4.3.4.1 TF-IDF

We use this technique to identify specific words with important sentiment-related information in Algerian dialect texts. This is achieved by considering their frequency in the sen-

tence and rarity in the entire dataset. Unlike Count Vectorizer, which only considers the number of word appearances, this technique provides a more nuanced approach. Table 4.3 below illustrates the difference between the two techniques.

	بزاف	ميش	مليحة
مليحه بزاف	1	0	1
ميش مليحة	1	1	0

	بزاف	ميش	مليحة
مليحة بزاف	0.579739	0.000000	0.814802
ميش مليحة	0.579739	0.814802	0.000000

Table 4.3: Comparison between count vectorizer and TF-IDF Vectorizer

To make the TF-IDF vectorizer works more efficiently with Latin written words and derived words, we need to preprocess our dataset with the three levels of preprocessing (Data cleaning, the trait of the Latin letters, and stemming).

4.3.4.2 Word2Vec

Word2Vec is a popular word embedding technique in natural language processing. It aims to learn distributed representations of words based on their contextual usage in a given text corpus. The basic concept is that words with close meanings or appearing in similar contexts will have similar vector representations.

Figure 4.4 shows the closest terms in the word2vec vocabulary after we train it on our corpus. In our case, the most similar words to education ,**التعليم**

Word	Similarity Score
التدريس	0.9126098
المدرسة	0.8856790
الطلاب	0.8554326
المناهج	0.8376532
التعلم	0.8209845
الجامعة	0.8005664
المعلمين	0.7967548
التدريب	0.7712345
الفصول	0.7667899
التربية	0.7564210

Table 4.4: Top similar words using Word2Vec model.

4.3.4.3 FastText

FastText is considered as an improvement over Word2Vec in specific scenarios due to its ability to consider sub-word information. This feature enables FastText to handle out-of-vocabulary words and capture morphological variations more effectively. FastText is a preferred choice for sentiment analysis in Algerian dialects, where dialect-specific words and morphological variations can pose challenges for Word2Vec. Due to their characteristics, we do not need to perform stemming on the training corpus or the classification data when we work with FastText as word representation.

Figure 4.5 shows the closest terms in the FastText vocabulary after we train it on our corpus. In our case, the most similar words to a lot **بزاف**

Word	Similarity Score
بزاف	0.94256
بالبزاف	0.86506
بزاااااف	0.84788
بززززاف	0.84765

Table 4.5: Top similar words using FastTest model

4.4 Training

This section will explore our model construction approach with three different categories of algorithms: Machine Learning (ML), Deep Learning (DL), and pre-trained transformer models.

- **Machine Learning:** In our work, we retained all sentiment categories, including the neutral class, to fully capture the diversity of sentiments expressed in the Algerian dialect. The dataset was split into 80% for training and 20% for testing. For feature extraction, we experimented with several techniques such as TF-IDF, Word2Vec, and FastText to convert the textual data into numerical representations. These features were then used to train two classical machine learning classifiers: Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB). Keeping the neutral class allowed the models to learn the fine distinctions between neutral, positive, and negative sentiments, resulting in a more comprehensive classification performance across all classes.

- **Deep Learning:** In this work, we divided our dataset into three partitions: 60% for training, 20% for validation, and 20% for testing. The text data was converted into numerical representations using one of the feature extraction techniques (TF-IDF, Word2Vec, or FastText).

  We employed a deep neural network architecture that includes a Bidirectional LSTM layer, which allows the model to capture both forward and backward dependencies within the sequences. This structure helps the model better understand the contextual relationships in Algerian dialect sentences.

  We applied one-hot encoding to the sentiment labels, representing each sentiment class (positive, neutral, and negative) as a binary vector. The model was trained using the deep neural network with a SoftMax activation function in the output layer to perform multi-class classification.

  For model optimization, the Hyperband technique was used to explore and select the best hyperparameters such as batch size and dropout rate to maximize performance. Additionally, the early stopping technique was applied to monitor the validation accuracy and prevent overfitting by stopping the training when the model ceased to improve.

- **Pretrained transformer modules:** To fine-tune the transformer models for the sentiment analysis task, we selected the most popular pre-trained transformer models available on the Hugging Face Hub that support the Arabic language and Algerian dialect.

  Initially, the dataset was divided into 60% for training, 20% for validation, and 20% for testing. The text data was tokenized using the pre-trained *AutoTokenizer* corresponding to each model. After tokenization, the models were fine-tuned using *AutoModel*. Finally, the trained models were evaluated on the test partition.

  In our work, we fine-tuned the following transformer models:

- **BERT Multilingual base model**: pre-trained with Wikipedia data covering 104 languages, including Arabic and French.

- **DziriBERT**: pre-trainedspecifically on Algerian dialect using 1 million tweets.
- **AraBERT** : pre-trained on 200 million Arabic sentences.
- **MarBERT** : pre-trained on the Moroccan dialect, which is linguistically close to the Algerian dialect.
- **XLM-RoBERTa base**: developed by META (formerly Facebook), pre-trained on 2.5 TB of multilingual data covering 100 languages including Arabic and French

4.5 Evaluation metrics

We used four evaluation metrics: Accuracy, F1 score, Precision, Recall.

Because we work on a balanced dataset, the primary metric we employed during evaluation and experimentation is Accuracy that represents the total number of words by the number of correctly transliterated words.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Moreover, we evaluate our model with an F1 score on the three categories (negative, neutral, and positive), which is given by the equation:

$$F1 = \frac{2 * P * R}{P + R} \tag{4.2}$$

Where the Precision represents the percentage of true-positive predictions out of all positive predictions or, in a simple way, represents the number of sentiments correctly labeled as belonging to the positive class divided by the total number of sentiments labeled as belonging to the positive class.

$$P = \frac{TP}{TP + FP} \tag{4.3}$$

Conclusion

Throughout this chapter, a complete practical framework has been established for data preparation and model construction for Algerian dialect sentiment classification. The work encompassed merging multi-source datasets, performing advanced linguistic preprocessing tailored to the dialect's characteristics, and employing diverse feature extraction techniques. Various models were trained using machine learning algorithms, deep learning architectures, and state-of-the-art pre-trained transformer models. A thorough evaluation using accuracy, recall, precision, and F1-score metrics was conducted to analyze and compare model performances. This systematic approach enabled the development of models capable of effectively addressing the complexities of the Algerian dialect.

Chapter 5: Implementation et Experiments

Introduction

In this chapter, we present the results of the experiments conducted on a set of machine learning models, deep learning models, and pre-trained transformer models, with the aim of analyzing their performance in the task of sentiment classification for the Algerian dialect. The models were evaluated using various metrics such as precision, recall, and F1-score, with a detailed analysis of the results for each model presented through tables and charts. In addition, we tested the models on real-life examples to assess their generalization ability and their performance on real-world sentences outside of the training data.

5.1 Materials

"We do all our work and training on the Google Colab environment for its user-friendly interface and rich set of programming tools. Our configuration for the environment included the following specifications:"

CPU	**Intel Xeon CPU @2.20 GHz**
RAM	13 GB RAM
GPU	Tesla K80
VRAM	12 GB GDDR5

Table 5.1: The performance of the used material

5.1.1 Tools

"We opted for the Python programming language, renowned for its capabilities in artificial intelligence and data manipulation. The libraries employed during the programming process are detailed in Table 5.2"

Library	**Description**
Pandas	an open-source data analysis and manipulation tool in Python. It is used for reading datasets and performing data analysis tasks.
Scikit-learn (sklearn)	a library for machine learning, includes classification, regression, clustering, and dimensionality reduction.
NumPy	an important library for scientific computing. It includes a multidimensional array object and derived objects such as masked arrays and matrices.
(RE) Regular Expression	package that provides tools for working with regular expressions, That are a sequence of characters with a specific syntax for pattern searching and string matching.
Matplotlib	a Python package that provides functionalities for creating static, animated, and interactive visualizations.
Camel Tools [36]	a software package that provides various language processing and analysis tools specifically designed for the Arabic language, we use its dialect detection function
PyTorch	deep learning framework that offers a flexible and efficient platform for developing and training neural networks.
langdetect	a library that identify the language of a text or document
Deep-translator	a library for translation tasks using machine translation services such as Google Translate
Tasphayne	a library that provides useful utilities for Arabic text processing and NLP tasks, including tokenization, normalization, stemming, and more.
NLTK	a popular library that provides various tools for NLP, we use its Arabic stopwords
Aaransia	a library that transliterates text from Arabizi to Arabic
Keras	a high-level neural network library that provides a user- friendly interface for building and training deep learning models on top of other frameworks like.
Gensim	is a library for topic modeling and document similarity analysis, providing efficient tools for working with large text corpora and building word embeddings.
Transformers [51]	It is a powerful library that offers state-of-the-art natural language processing (NLP) capabilities including pre-trained models for tasks like text classification.

Table 5.2: The used python libraries

5.2 Coding

5.2.1 Machine Learning

TF-IDF

After applying full text preprocessing — which included normalization, cleaning diacritics, reducing letter repetitions, removing non-Arabic characters, and filtering out Arabic stopwords — the final cleaned dataset was used to fit a TF-IDF vectorizer. The TF-IDF vectorizer extracted **15,132 unique terms** from the dataset. The dataset was then split into training and testing subsets. Both subsets were transformed into feature vectors using the TF-IDF matrix. Two traditional classifiers were trained on these features:

- **Support Vector Machine (SVM - SVC):** using default linear kernel.
- **Multinomial Naïve Bayes (MNB).**

Word2Vec and FastText

The Word2Vec model was trained directly on the cleaned Algerian dialect corpus using Gensim's Word2Vec implementation, with a vector size of **300 dimensions** and a minimum word frequency threshold of 2 (`min_count=2`). The resulting Word2Vec vocabulary contained approximately **787,389 unique tokens**.

Similarly, FastText was trained on the same corpus using Gensim's FastText implementation, producing word embeddings of **300 dimensions** while capturing subword n-gram representations for better handling of out-of-vocabulary words. The FastText vocabulary contained around **338,664 unique words** with their corresponding subword representations.

For both embeddings, the sentiment dataset was transformed by averaging the word embeddings of each sentence to obtain a fixed-length feature vector, which was used to train both SVM and Naïve Bayes classifiers.

5.2.2 Deep Learning

The cleaned dataset was divided into three partitions:

- 60% for training,
- 20% for validation,
- 20% for testing.

The textual data was tokenized using **Keras Tokenizer**, which assigned unique integer indices to each word in the vocabulary. After tokenization, sentences were converted into sequences of integers.

Since neural networks require fixed-length input sequences, padding was applied using Keras' `pad_sequences()` to ensure all sequences reached the same maximum length, which was computed as the 90th percentile of sequence lengths.

For embedding initialization:

- Either random embeddings (learned from scratch),
- Or pretrained Word2Vec embeddings,
- Or pretrained FastText embeddings were loaded into the embedding layer.

The architecture used in deep learning models consisted of:

- **Embedding layer:** initialized with either pretrained embeddings or randomly initialized weights.
- **Bidirectional LSTM layer:** with an output size of 256 units to capture both forward and backward dependencies.
- **GlobalMaxPool1D layer:** for selecting the most important features across the sequence.
- **Three Dense layers:** with output sizes of 128, 64, and 32 units, each followed by Dropout layers for regularization.
- **Output layer:**
  - For multi-class classification: Dense layer with 3 neurons and Softmax activation.

The models were compiled using **Adam optimizer** and trained using cross-entropy loss with class balancing applied to handle class imbalance.

5.2.3 Transformer Models

In this phase, we fine-tuned several pre-trained transformer models on the Algerian sentiment analysis task. The models include:

- **BERT Multilingual (bert-base-multilingual-cased)**
- **DziriBERT (alger-ia/dziribert)**
- **AraBERT (aubmindlab/bert-base-arabertv2)**
- **MarBERT (UBC-NLP/MARBERTv2)**
- **XLM-RoBERTa (xlm-roberta-base)**

For fine-tuning, we used **Huggingface's Transformers** library. The dataset was tokenized using each model's corresponding tokenizer and fine-tuned for **3 epochs** using the AdamW optimizer with a learning rate of `2e-5`.

The pre-trained models demonstrated superior performance due to their ability to capture contextualized word representations, especially DziriBERT, which benefited from pretraining on Algerian Arabic dialect data.

5.3 Results

models	Precision	Recall	F1	Accuracy
TF-IDF, SVM	65%	66%	67 %	70,8%
TF-IDF, NB	67%	68%	68%	68.9%
Word2vec,Transformer	**75%**	**75%**	**75%**	**75.6%**
FastText,Transformer	**76%**	**77%**	**78%**	**78,9%**
Tokenizer, BiLSTM	71%	71%	73%	71%
Word2vec, BiLSTM	57%	57%	57%	58%
FastText, BiLSTM	61 %	61 %	61 %	63%
BERT multilanguage	75%	75%	75%	75,1%
DziriBERT	**82%**	**82,6%**	**82%**	**82,2%**
AraBERT	80%	80%	80%	80%
MarBERT	80%	81%	81,2%	81,6%
XLM RoBERTa	80%	80%	80%	80%

Table 5.3: The results of our models.

Evaluation Results of the Proposed Model for Algerian Dialect Sentiment Classification

After training the updated model on the Algerian dialect sentiment data, it demonstrated stable and effective performance. As shown in Figure 5.1 , the training loss decreased from 0.85 to 0.07, while the validation loss started increasing after epoch 5. The training accuracy increased from 60% to 96%, while validation accuracy stabilized at 78%.

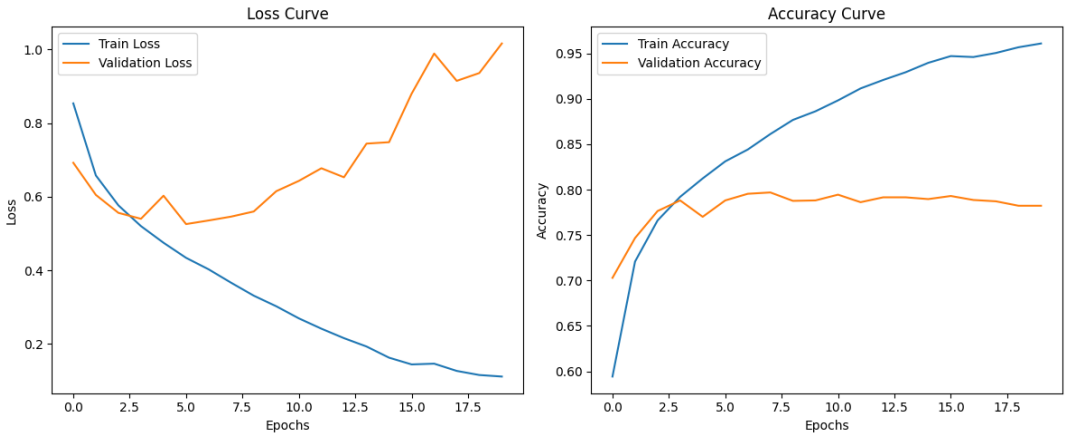

Figure 5.1: Training and validation loss and accuracy curves of the proposed model .

The confusion matrix in Figure 5.2 shows that the model correctly classified 531 negative,

453 neutral, and 610 positive samples. These results confirm the model's effectiveness in handling Algerian dialect sentiment data.

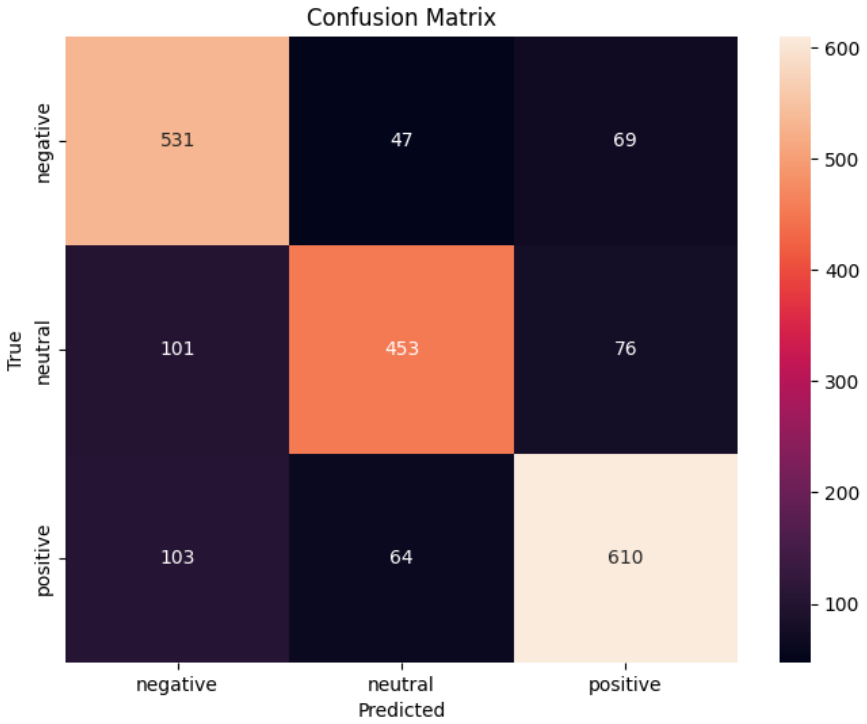

Figure 5.2: Confusion matrix of the proposed model .

5.3.1 Results Discussion for Machine Learning Models

In our experiments, the traditional machine learning algorithms displayed varying performances depending on the feature extraction technique used.

When applying Naïve Bayes, it achieved an accuracy of 68% with TF-IDF, which slightly decreased when using word embeddings, scoring 67% with Word2Vec and 68% with FastText. This demonstrates the typical behavior of Naïve Bayes, which assumes feature independence, making it better suited for sparse vectorization methods like TF-IDF where word presence or absence holds more discriminative power.

On the other hand, SVM consistently outperformed Naïve Bayes across all embeddings. It achieved:

70% accuracy with TF-IDF,

75% accuracy with Word2Vec,

and 78% accuracy with FastText.

This confirms the known strength of SVM in handling high-dimensional data, capturing more complex boundaries between classes compared to Naïve Bayes. The results also reinforce the finding that TF-IDF remains a very strong feature extraction method for classic machine learning, especially when dealing with domain-specific vocabularies, as it focuses on term frequency importance rather than purely contextual word similarities

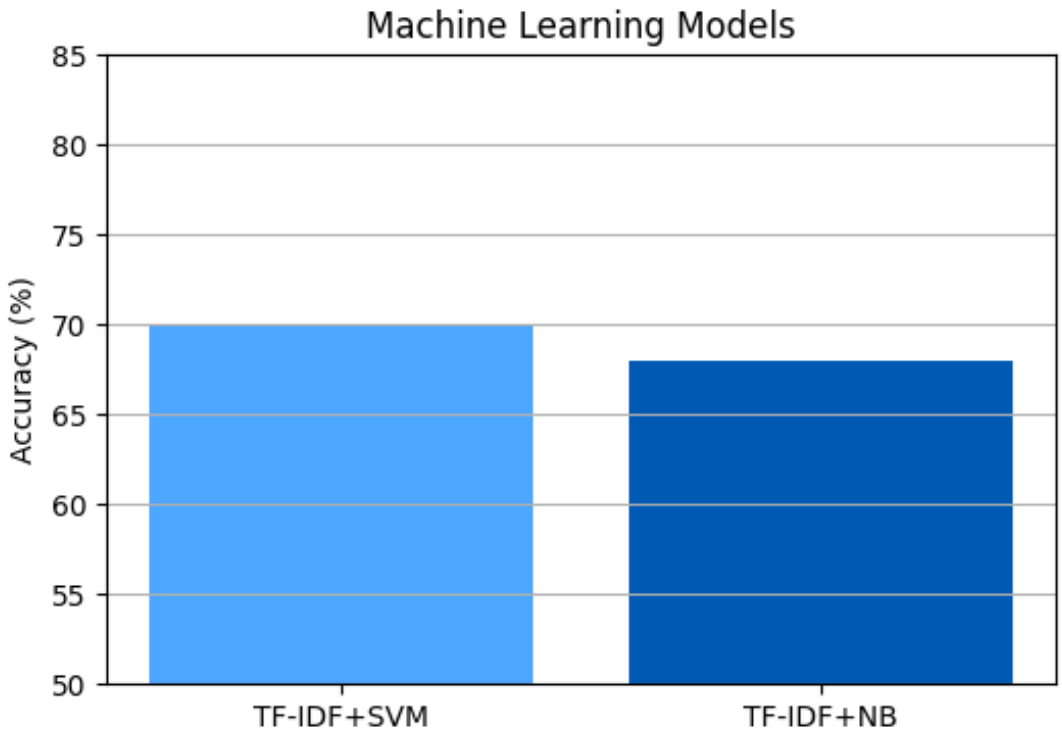

Figure 5.3: Bar chart comparison between different ML models .

5.3.1.1 Results Discussion for Deep Learning Models

In the case of deep learning models, we experimented with various architectures including BiLSTM combined with different word representations. When using a simple tokenizer-based approach (without pretrained embeddings), BiLSTM achieved an accuracy of 71%.

Upon integrating pretrained embeddings:

The combination of Word2Vec with BiLSTM resulted in a performance decrease, yielding 58% accuracy.

The FastText embeddings with BiLSTM improved the performance slightly to 63% accuracy.

These results suggest that while contextual embeddings like FastText can help deep models better capture subword information, they still face challenges in fully modeling the nuances of the Algerian dialect, especially when compared to transformer-based approaches. Overall, the BiLSTM models demonstrated moderate performance, but clearly benefited more when combined with pretrained embeddings.

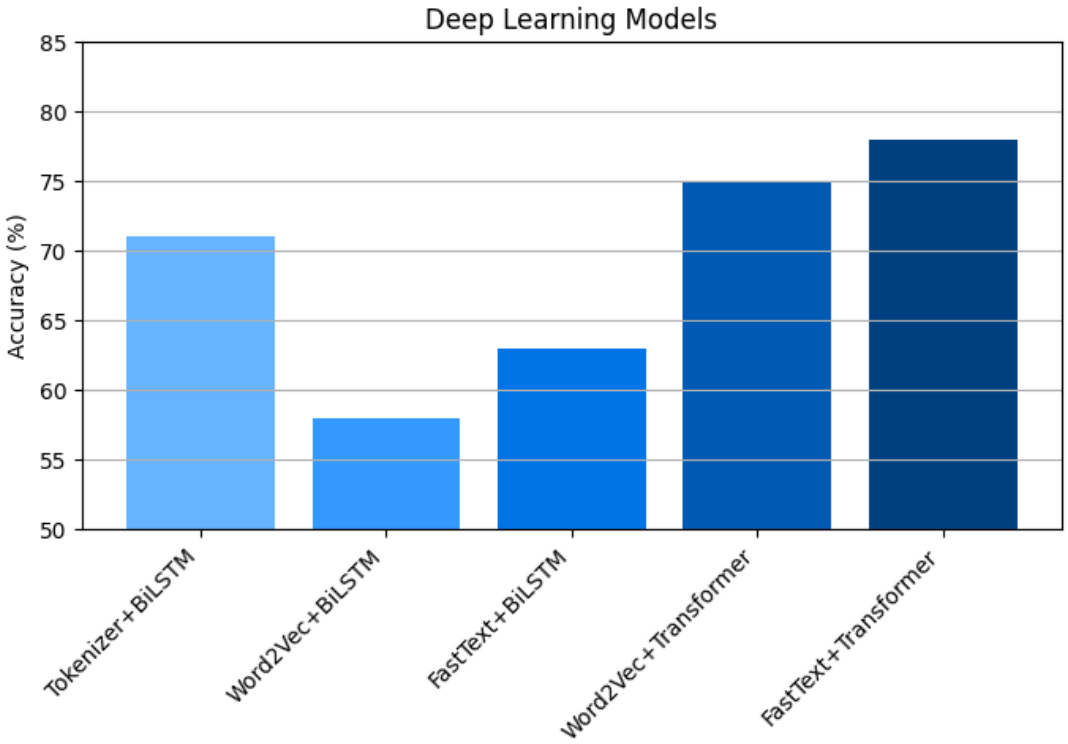

Figure 5.4: Bar chart comparison between different DL models

5.3.1.2 Results Discussion for Transformer-Based Models

The best performances were obtained using pretrained transformer-based models, which significantly outperformed both machine learning and traditional deep learning architectures.

After fine-tuning multiple pretrained models on our Algerian sentiment dataset, the results were as follows:

DziriBERT achieved the highest accuracy of 82%, outperforming all other models thanks to its pretraining specifically on Algerian dialect data.

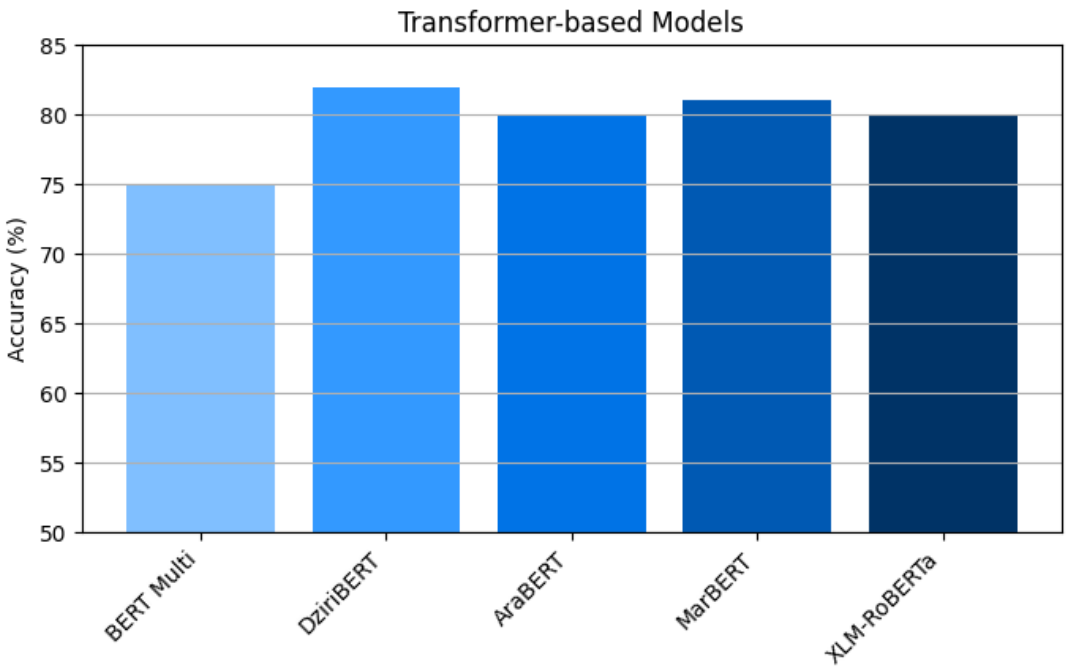

Figure 5.5: Bar chart comparison between different pre-trained transformer models.

MarBERT reached 81% accuracy, benefiting from its pretraining on the Moroccan dialect which shares many similarities with Algerian Arabic.

AraBERT and XLM-RoBERTa both achieved 80% accuracy, demonstrating strong performance despite being trained primarily on Modern Standard Arabic or multilingual corpora.

BERT Multilingual achieved 75% accuracy, slightly lower due to its more general multilingual nature.

The superior performance of transformers can be attributed to their self-attention mechanisms that effectively capture long-range dependencies and contextual relationships within the sentence, as well as their large-scale pretraining on diverse corpora, which enables robust language understanding. Finally, the machine learning results varied from 67% to 74 %. The results are illustrated in

The results clearly demonstrate the superiority of transformer-based models attributed to their attention mechanisms, effectively capturing long-range dependencies and contextualized representations (as explained in the previous chapter). The language understanding capabilities of the models are further improved through pre-training on extensive corpora

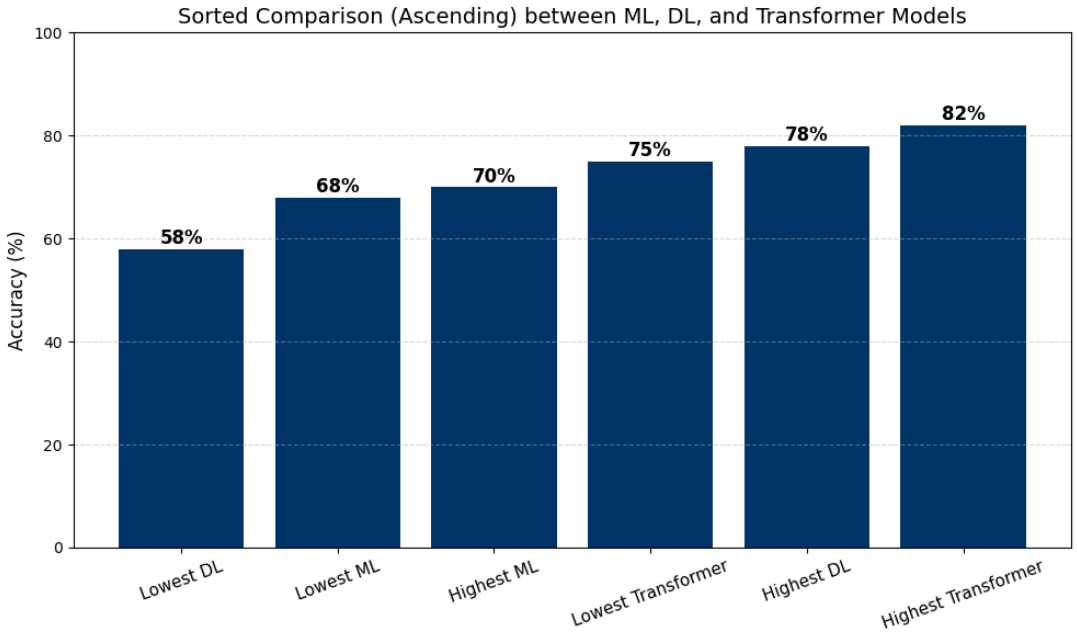

Figure 5.6: Bar chart comparison between different techniques.

and architectural advancements, resulting in higher performance in our work. To do a detailed analysis of the results and performance of our highest-scoring model DziriBERT, we will present the classification results obtained from the report analysis. The results are summarized in Table 5.3, highlighting the precision, recall, and F1-score of each label and of the average of them. This comprehensive evaluation offers valuable insights into the performance of the model and its ability to accurately classify sentences in the three different categories. Additionally, we will present the confusion matrix for our sentiment

Label	Precision	Recall	F1-Score
Positive	87%	83%	85%
Neutral	80%	82%	81%
Negative	78%	80%	79%
Score	**82%**	**82%**	**82%**

Table 5.3: Classification report of DziriBERT

analysis task (Figure 5.5), which involves classifying text into three sentiment categories: Negative, Neutral, and Positive. The confusion matrix provides a visual representation of how well our classification models performed in predicting the correct sentiment labels. By examining the confusion matrix, we can gain insights into the distribution of predictions across the sentiment categories and assess the accuracy of the models in correctly identifying the sentiments expressed in the text. This analysis allows us to evaluate the strengths and weaknesses of our sentiment analysis approach and get conclusions about

the performance of the models in handling the three labeled sentiment classification tasks.

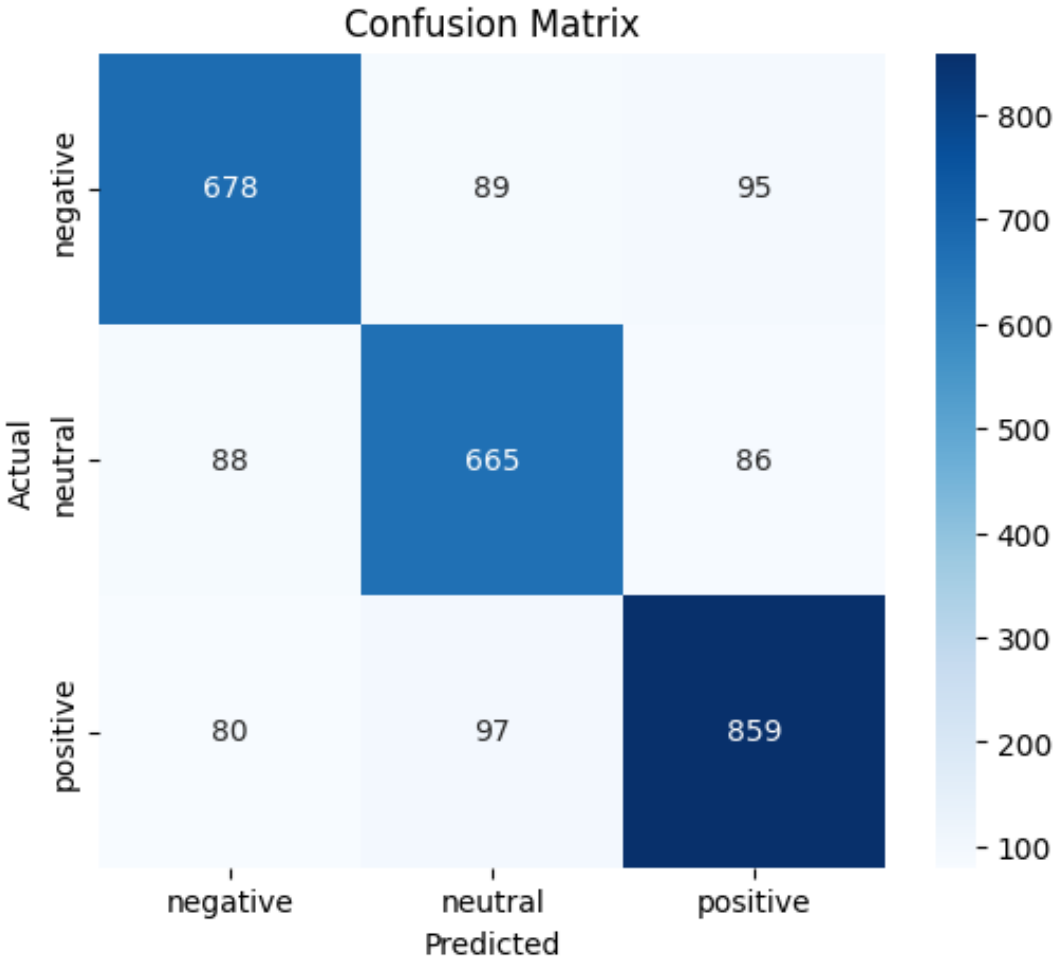

Figure 5.7: Confusion matrix of DziriBERT model.

Finally, we will present some examples of model prediction. The first example, مليح ؟ واش تنصحني ؟ تيليفون This sentence reflects a neutral inquiry where the user is seeking information without explicitly expressing positive or negative sentiment. The model correctly classified it as neutral. However, when the question marks are removed, and the sentence becomes:واش تنصحني التليفون مليح the model misclassified it as positive, indicating that the presence of question marks helps the model better differentiate between questions and positive affirmations.

In another example, the sentence تليفون مليح بصح غالي carries both positive and negative sentiments. The model classified this sentence as positive, while the true label is neutral due to the mixed sentiment expressed. This illustrates the challenge faced by the model

in handling sentences with conflicting opinions.

These examples provide valuable insights into the model's behavior and emphasize the importance of preserving key linguistic cues, such as punctuation, during the preprocessing phase to improve classification performance.

Sentence (in Arabic)	True Label	Predicted Label
واش تنصحني ؟ تيليفون مليح ؟	Neutral	Neutral
واش تنصحني التليفون مليح	Neutral	Positive
تليفون مليح بصح غالي	Neutral (Mixed)	Positive

Table 5.4: Examples of classification with our highest-scoring model.

Conclusion

To summarize, this chapter demonstrates a thorough evaluation of various sentiment analysis models applied to the Algerian dialect. The experimental results confirm the superiority of transformer-based models, particularly DziriBERT, which achieved the highest accuracy due to its specific pretraining on Algerian Arabic. While classical machine learning algorithms and deep learning models showed acceptable performances, their limitations were evident compared to the rich contextual capabilities of transformers. The detailed analysis, confusion matrices, and real-world examples further highlight the strengths and challenges of each approach, offering valuable insights for future work in sentiment analysis for low-resource dialects like Algerian Arabic.

Conclusion

In conclusion, this thesis addressed the task of sentiment analysis in the Algerian dialect, recognizing the need to develop specialized models and techniques capable of accurately analyzing and understanding sentiments expressed within this particular dialectal context.

Our research followed several key stages. Initially, we collected and manually annotated a dataset consisting of 15,000 comments. The data then underwent preprocessing, taking into account the unique linguistic features of the Algerian dialect.

Throughout this work, we applied various techniques to extract numerical features from the textual data, such as TF-IDF, Word2Vec, FastText, and tokenization. These techniques allowed us to better represent the linguistic nuances and dialect-specific elements present in the Algerian dialect.

Subsequently, we employed both traditional machine learning algorithms, including SVM and Naïve Bayes, as well as deep learning techniques, particularly Long Short-Term Memory (LSTM), to train and develop models capable of accurately predicting sentiment. Our results demonstrated the effectiveness of deep learning models in capturing complex relationships and patterns within the data.

Furthermore, we recognized the importance of pre-trained transformer-based models in sentiment analysis. By fine-tuning models such as DziriBERT on Algerian dialect data, we achieved advanced performance in sentiment analysis tasks.

Overall, this research contributes to the field of sentiment analysis by providing valuable insights into the sentiments, opinions, and attitudes expressed by Algerian speakers. By analyzing sentiment in this dialectal context, it becomes possible to gain a deeper understanding of public opinion, support better decision-making processes, and enhance

various applications such as social media monitoring and market research targeting the Algerian community.

Future Perspectives

Although this research project has yielded significant results, there remains considerable room for further development. Future work may include increasing the size of the dataset, introducing a new ``mixed sentiment'' classification category, and exploring additional preprocessing techniques such as stemming.

These improvements would contribute to a more comprehensive understanding of sentiment analysis in the Algerian dialect and further enhance the models' efficiency and adaptability in handling the specific linguistic features of this dialect.

References

[1] S. Harrat, K. Meftouh, M. Abbas, and K. Smaili, ``Building resources for algerian arabic dialects," *International Journal of Computational Linguistics*, vol. 1, no. 1, pp. 15--25, 2014, available at: slmhrrt@gmail.com, Karima.meftouh@univ-annaba.org, m_abbas04@yahoo.fr.

[2] S. A. Amrani, M. Lazaar, and K. E. E. Kadiri, ``Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Computer Science*, vol. 127, pp. 511--520, 2018.

[3] H. Saadane and N. Habash, ``A conventional orthography for algerian arabic," in *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Grenoble, France, 2015.

[4] P. Balaji, O. Nagaraju, and D. Haritha, ``Levels of sentiment analysis and its challenges: A literature review," in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE, 2017, pp. 436--439.

[5] A. Kumar and T. M. Sebastian, ``Sentiment analysis: A perspective on its past, present and future," *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 4, no. 10, pp. 1--14, 2012. [Online]. Available: http://www.mecs-press.org/ijisa/ijisa-v4-n10/v4n10-1.html

[6] Shelf, ``Building smarter ai with data augmentation," https://shelf.io/blog/data-augmentation/, n.d., accessed: 2025-0.

[7] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Munoz-Merino, I. Estevez-Ayres, and C. D. Kloos, ``Sentiment analysis in moocs: A case study," in *2018 IEEE Global*

Engineering Education Conference (EDUCON). IEEE, 2018, pp. 1489--1496. [Online]. Available: https://ieeexplore.ieee.org/document/8363409/

[8] R. B. Hubert, E. Estevez, A. Maguitman, and T. Janowski, ``Analyzing and visualizing government-citizen interactions on twitter to support public policy-making," *ACM Journal of Data and Information Quality (JDIQ)*, vol. 12, no. 2, pp. 1--20, 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3360001

[9] R. M. Saeed, S. Rady, and T. F. Gharib, ``An ensemble approach for spam detection in arabic opinion texts," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 1, pp. 1407--1416, 2022.

[10] B. Liu, *Sentiment Analysis and Opinion Mining*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012, vol. 5, no. 1.

[11] D. Maynard and M. A. Greenwood, ``Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavík, Iceland: ELRA, 2014, pp. 4238--4243.

[12] M. Ahmed, Q. Chen, and Z. Li, ``Constructing domain-dependent sentiment dictionary for sentiment analysis," *Neural Computing and Applications*, vol. 32, pp. 14 719--14 732, 2020.

[13] M. Mataoui, O. Zelmati, and M. Boumechache, ``A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic," *Research in Computing Science*, vol. 110, no. 1, pp. 55--70, 2016.

[14] I. Guellil, A. Adeel, F. Azouaou, and A. Hussain, ``Sentialg: Automated corpus annotation for algerian sentiment analysis," in *Advances in Brain Inspired Cognitive Systems: 9th International Conference, BICS 2018*, ser. Proceedings. Springer, 2018, pp. 557--567.

[15] A. Abdelli, F. Guerrouf, O. Tibermacine, and B. Abdelli, ``Sentiment analysis of arabic algerian dialect using a supervised method," in *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*. IEEE,

2019, pp. 1--6.

[16] L. Moudjari, K. Akli-Astouati, and F. Benamara, ``An algerian corpus and an annotation platform for opinion and emotion analysis,'' in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 1202--1210.

[17] L. Ouchene and S. Bessou, ``Sentiment analysis for algerian dialect tweets,'' in *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2022, pp. 235--239.

[18] 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, pp. 897--903, 2022.

[19] A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, ``Dziribert: a pre-trained language model for the algerian dialect,'' *arXiv preprint arXiv:2109.12346*, 2021.

[20] M. Laifa and D. Mohdeb, ``Sentiment analysis of the algerian social movement inception,'' *Data Technologies and Applications*, 2023, ahead-of-print.

[21] NVIDIA, ``Fundamentals of deep learning,'' https://www.nvidia.com/content/dam/en-zz/Solutions/deep-learning/, 2025, accessed: 2025-05-12.

[22] B. Mahesh *et al.*, ``Machine learning algorithms—a review,'' *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381--386, 2020, [Internet].

[23] A. Graves and J. Schmidhuber, ``Framewise phoneme classification with bidirectional lstm and other neural network architectures,'' *Neural Networks*, vol. 18, no. 5-6, pp. 602--610, 2005.

[24] Z. C. Lipton, J. Berkowitz, and C. Elkan, ``A critical review of recurrent neural networks for sequence learning,'' *arXiv preprint arXiv:1506.00019*, 2015, [Online; accessed 2025-05-17].

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ``Gradient-based learning applied to document recognition,'' *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278--2324, 1998.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ``Attention is all you need,'' in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ``Bert: Pre-training of deep bidirectional transformers for language understanding,'' *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[28] OpenAI, ``Gpt-4 technical report,'' *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, ``Roberta: A robustly optimized bert pretraining approach,'' *arXiv preprint arXiv:1907.11692*, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[30] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, ``Transformer-xl: Attentive language models beyond a fixed-length context,'' *arXiv preprint arXiv:1901.02860*, 2019. [Online]. Available: https://arxiv.org/abs/1901.02860

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ``Bert: Pre-training of deep bidirectional transformers for language understanding,'' *arXiv preprint arXiv:1810.04805*, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[32] W. Antoun, F. Baly, and H. Hajj, ``Arabert: Transformer-based model for arabic language understanding,'' *arXiv preprint arXiv:2003.00104*, 2020. [Online]. Available: https://arxiv.org/abs/2003.00104

[33] A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, ``Dziribert: A pre-trained language model for the algerian dialect,'' *arXiv preprint arXiv:2109.12346*, 2021. [Online]. Available: https://arxiv.org/abs/2109.12346

[34] Y. LeCun, Y. Bengio, and G. Hinton, ``Deep learning,'' *Nature*, vol. 521, no. 7553, pp. 436--444, 2015.

[35] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[36] T. Xu and F. Liang, ``Machine learning for hydrologic sciences: An introductory overview," *WIREs Water*, vol. 8, no. 5, p. e1533, 2021.

[37] S. Brown, ``Machine learning, explained," https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained, 2021, mIT Sloan School of Management, April 2021. [Online; accessed 2025-06-05].

[38] M. T. Almuqati, F. Sidi, S. N. Mohd Rum, M. Zolkepli, and I. Ishak, ``Challenges in supervised and unsupervised learning: A comprehensive overview," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 14, no. 4, pp. 1449--1455, 2024.

[39] B. Mahesh, ``Machine learning algorithms—a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381--386, 2020, [Internet].

[40] ------, ``Machine learning algorithms—a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381--386, 2020, [Internet].

[41] L. Breiman, ``Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5--32, 2001.

[42] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, ``Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1--16, 2019.

[43] ``A comparative study of machine learning algorithms," Master's thesis, McMaster University, Hamilton, Ontario, Canada, 2018, [Online]. Available: https://macsphere.mcmaster.ca/handle/11375/23649.

[44] M. Salhi, S. Benali, and M. Zegadi, ``Sentiment analysis of algerian dialect on twitter: The case of djezzy telecom company," *Journal of Information Science and Engineering*, vol. 37, no. 1, pp. 123--135, 2021.

[45] A. Chader, D. Lanasri, L. Hamdad, M. C. E. Belkheir, and W. Hennoune, ``Sentiment analysis for arabizi: Application to algerian dialect," in *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART)*, vol. 2, 2019, pp. 475--482.

[46] M. Mataoui, A. Guessoum, and H. Boubekeur, ``Twifil: A manually annotated twit-

ter dataset for sentiment analysis in algerian dialect," *International Journal of Computational Linguistics*, vol. 8, no. 4, pp. 50--62, 2020.

[47] ------, ``Sentiment analysis dataset for algerian dialect in social media," *Arabian Journal of Science and Engineering*, vol. 45, no. 5, pp. 3671--3684, 2020.

[48] Y. Name, ``Sentiment analysis dataset for algerian automotive domain," Unpublished dataset, 2025.

2024/2025