الجمهورية الشعبية الديمقراطية الجزائرية

PEOPLE's DEMOCRATIC REPUBLIC of ALGERIA

Ministry of Higher Education and Scientific Research

University of Ferhat Abbas Setif 1

Faculty of Sciences

Department of Computer Science



PhD Dissertation

Presented by:

Lamia OUCHENE

As a Requirement for Obtaining the **Doctoral Degree** in Computer Science

Option: Software Engineering and Human-Machine Interface

Enhancing Human-Machine Interaction through Sentiment Analysis of the Algerian Dialect

---

PhD Defense Board:

Lekhfif Abdelaziz	Pr. at Univ. of Ferhat Abbas Setif 1	President
Bessou Sadik	Pr. at Univ. of Ferhat Abbas Setif 1	Supervisor
Toumi Lyazid	Pr. at Univ. of Ferhat Abbas Setif 1	Examiner
Akrouf Samir	Pr. at Univ. of Mohamed Boudiaf M'sila	Examiner
Baghoura Mohamed Amine	MCA. at Univ. of BBA	Examiner
Touharia Mohamed	Pr. Retired Professor	Invited
Boubetra Abdelhak	Pr. Retired Professor	Invited

Defended on: 01/ 06/ 2026

Abstract

This thesis investigates sentiment and subjectivity analysis in the Algerian Arabic dialect (Darija) with the overarching aim of enhancing human–machine interaction for dialect speakers. Algerian Arabic presents significant linguistic challenges, including non-standardized spelling, extensive code-switching, and strong regional variation, which complicate natural language processing tasks and contribute to the scarcity of reliable computational resources.

To address these limitations, this work introduces two major contributions. First, we develop and publicly release AlgVec, a suite of word embedding models trained on over 32 million tokens extracted from user-generated content on X (formerly Twitter), covering both Algerian Arabic script and Arabizi. These embeddings are trained using Word2Vec (Skip-gram and CBOW) and FastText architectures and evaluated through intrinsic tasks such as word similarity and nearest-neighbor retrieval, as well as a dialect-focused benchmark adapted from DiaLex. Second, we construct a publicly available corpus of 18,772 manually annotated Algerian dialect tweets labeled into three sentiment classes: positive, negative, and neutral. A dedicated preprocessing pipeline is proposed to handle dialect-specific characteristics, including emoji usage, Arabizi, and code-switching.

We conduct a comprehensive benchmark of traditional machine learning approaches (Support Vector Machines and Naive Bayes), deep neural models (Long Short-Term Memory networks and Convolutional Neural Networks), and transformer-based architectures. In addition, we explore a combined embedding strategy that integrates AlgVec with MARBERT for downstream sentiment classification. Experimental results demonstrate that transformer-based models, particularly the MARBERT–LSTM configuration, achieve state-of-the-art performance with an accuracy of 91.23%, significantly outperforming traditional methods and widely used Modern Standard Arabic embeddings on dialectal tasks.

These findings highlight the importance of dialect-aware sentiment analysis for developing more intuitive and culturally aware human–machine interfaces, enabling emotionally intelligent systems such as virtual assistants, customer service platforms, and interactive educational tools to better understand and respond to users' emotional states expressed in their native dialect.

الملخص

تتناول هذه الأطروحة تحليل المشاعر والذاتية في اللهجة العربية الجزائرية (الدارجة)، بهدف تحسين التفاعل بين الإنسان والآلة لدى متحدثي اللهجات. تتميز اللهجة الجزائرية بتحديات لغوية معقدة، من بينها غياب التوحيد الإملائي، وكثرة التناوب اللغوي بين العربية واللغات الأجنبية، بالإضافة إلى التنوع الجهوي الواسع، وهو ما يجعل مهام المعالجة الآلية للغة الطبيعية أكثر صعوبة ويساهم في ندرة الموارد الحاسوبية الموثوقة.

لمعالجة هذه الإشكاليات، يقدم هذا العمل مساهمتين رئيسيتين. أولاً، نقترح وننشر AlgVec، وهي مجموعة من نماذج التمثيل الدلالي للكلمات تم تدريبها على أكثر من 32 مليون وحدة لغوية مستخرجة من محتوى المستخدمين على منصة X (تويتر سابقاً)، وتشمل النصوص المكتوبة باللهجة الجزائرية بالحرف العربي وبالArabizi. تم تدريب هذه النماذج باستخدام معماريات (Word2Vec (Skip-gram CBOW وFastText، وتقييمها من خلال مهام داخلية مثل قياس تشابه الكلمات واسترجاع أقرب الجيران، إضافة إلى اختبار معياري مخصص للهجات مقتبس من DiaLex.

ثانياً، قمنا ببناء مجموعة بيانات عامة تضم 18,772 تغريدة باللهجة الجزائرية، تم وسمها يدوياً وفق ثلاث فئات للمشاعر: إيجابي، سلبي، ومحايد. كما تم اقتراح خط معالجة مسبقة مخصص للتعامل مع الخصائص اللهجية، بما في ذلك استخدام الرموز التعبيرية، والArabizi، والنصوص المختلطة لغوياً.

أجرينا تقييماً شاملاً لمقاربات التعلم الآلي التقليدية، مثل آلة المتجهات الداعمة وNaive Bayes، ونماذج التعلم العميق، بما في ذلك شبكات الذاكرة طويلة قصيرة المدى (LSTM) والشبكات العصبية الالتفافية (CNN)، إضافة إلى النماذج المعتمدة على المحولات (Transformers). كما استكشفنا نهجاً هجينياً يجمع بين AlgVec وMARBERT في مهمة تصنيف المشاعر. أظهرت النتائج التجريبية أن النماذج المعتمدة على المحولات، وبخاصة نموذج MARBERT–LSTM، تحقق أداءً متقدماً بدقة بلغت 91.23%، متفوقةً بشكل ملحوظ على الأساليب التقليدية وتمثيلات الكلمات الشائعة المستخدمة في العربية الفصحى الحديثة عند التعامل مع النصوص اللهجية.

تؤكد هذه النتائج أهمية تحليل المشاعر المراعي للهجات في تطوير واجهات تفاعل إنسان--آلة أكثر ذكاءً وحساسية ثقافية، مما يتيح لأنظمة مثل المساعدات الافتراضية، وأدوات خدمة الزبائن، والمنصات التعليمية التفاعلية فهماً أفضل للحالات العاطفية للمستخدمين والاستجابة لها بلغتهم اليومية الطبيعية.

Résumé

Cette thèse étudie l'analyse du sentiment et de la subjectivité dans le dialecte arabe algérien (Darija) dans le but d'améliorer l'interaction homme–machine pour les locuteurs dialectaux. L'arabe algérien présente des défis linguistiques majeurs, notamment l'absence de normalisation orthographique, le code-switching fréquent et une forte variation régionale, ce qui complique les tâches de traitement automatique du langage naturel et contribue à la rareté des ressources computationnelles fiables.

Afin de répondre à ces limitations, ce travail propose deux contributions majeures. Premièrement, nous développons et mettons à disposition AlgVec, un ensemble de modèles de plongements lexicaux entraînés sur plus de 32 millions de tokens extraits de contenus générés par les utilisateurs sur X (anciennement Twitter), couvrant à la fois l'écriture arabe algérienne et l'Arabizi. Ces représentations sont entraînées à l'aide des architectures Word2Vec (Skip-gram et CBOW) et FastText, et évaluées à travers des tâches intrinsèques telles que la similarité lexicale et la recherche de plus proches voisins, ainsi qu'un benchmark dialectal adapté de DiaLex. Deuxièmement, nous construisons un corpus public de 18 772 tweets en dialecte algérien, annotés manuellement selon trois classes de sentiment : positif, négatif et neutre. Une chaîne de prétraitement dédiée est proposée pour gérer les spécificités dialectales, notamment l'utilisation des emojis, l'Arabizi et le texte en code-switching.

Nous réalisons une évaluation comparative approfondie de méthodes d'apprentissage automatique traditionnelles (machines à vecteurs de support et Naive Bayes), de modèles neuronaux profonds (réseaux LSTM et réseaux de neurones convolutifs), ainsi que d'architectures basées sur les transformeurs. Nous explorons également une approche hybride combinant AlgVec et MARBERT pour la classification du sentiment. Les résultats expérimentaux montrent que les modèles à base de transformeurs, en particulier la configuration MARBERT–LSTM, atteignent des performances de pointe avec une précision de 91,23%, surpassant significativement les approches traditionnelles et les embeddings couramment utilisés pour l'arabe standard moderne dans les tâches dialectales.

Ces résultats soulignent l'importance de l'analyse du sentiment sensible aux dialectes pour le développement d'interfaces homme–machine plus intuitives et culturellement adaptées, permettant à des systèmes intelligents tels que les assistants virtuels, les outils de service client et les plateformes éducatives interactives de mieux comprendre et répondre aux états émotionnels exprimés par les utilisateurs dans leur dialecte natif.

Authorship and Submission Declaration

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced. Where research has been carried out in collaboration with, or supported by others, this is duly acknowledged, and my contribution is indicated. This thesis has not been and will not be submitted, in whole or in part, to another university for the award of any other degree.

Signed: Lamia Ouchene
Date: June 01, 2026

© 2026 University of Ferhat Abbas 1

Acknowledgements

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ

First and foremost, I would like to express my sincere gratitude to الله عز وجل for giving me the strength, patience, and guidance throughout the journey of this thesis. As it is said in the Qur'an:

وَقُل رَّبِّ زِدْنِي عِلْمًا

"And say: My Lord, increase me in knowledge." (Qur'an, Surah Taha)

I would also like to thank my supervisor, **Dr. Sadik Bessou**, who supported me with his guidance, encouragement, and insightful discussions during my research.

Finally, my deepest gratitude goes to my beloved family for their unwavering love, patience, and sacrifices. Without their support, encouragement, and constant care, this work would not have been possible.

I dedicate this thesis to my mother, with all my love and gratitude.

Contents

1 Introduction . . . 1
- 1.1 Context and motivations . . . 1
- 1.2 Research Questions and Research Aims . . . 3
  - 1.2.1 Research questions . . . 3
  - 1.2.2 Research Aims . . . 3
- 1.3 Research Contributions . . . 4
- 1.4 Thesis plan . . . 5
- 1.5 Publications . . . 6
  - 1.5.1 International journals . . . 6
  - 1.5.2 International conferences and workshops . . . 6

2 Background . . . 8
- 2.1 Introduction . . . 8
- 2.2 Sentiment Analysis: An Overview . . . 8
  - 2.2.1 Levels of Sentiment Analysis . . . 9
  - 2.2.2 Techniques of Sentiment analysis . . . 10
- 2.3 The Role of X-Platform in Sentiment Analysis . . . 12
- 2.4 Arabic Language and Arabic dialects . . . 12
- 2.5 Algerian Dialect . . . 15
- 2.6 Algerian and Arabic Sentiment Analysis Challenges on Social Media . . . 16
  - 2.6.1 Language Challenges . . . 16
  - 2.6.2 X-Specific Challenges . . . 17
  - 2.6.3 Sentiment Analysis Challenges . . . 17
- 2.7 Human–Machine Interaction and Sentiment Interpretation . . . 17
- 2.8 Related Work . . . 19
  - 2.8.1 Related Work in Attention-Based Stopword Extraction . . . 19
  - 2.8.2 Related Work in AlgVec Word Embeddings . . . 20
  - 2.8.3 Related Work in Arabizi Sentiment Analysis . . . 21
  - 2.8.4 Related Work in Arabic Sentiment Analysis . . . 23
  - 2.8.5 Related Work in Human–Machine Interaction and Sentiment Analysis . . . 26

3 Attention-Based Stopword Extraction . . . 29
- 3.1 Introduction . . . 29
  - 3.1.1 Classical Stopword List Creation . . . 30
  - 3.1.2 Attention-based Stopword Extraction . . . 30
- 3.2 Experiments and Results . . . 31
  - 3.2.1 Stopword Removal Configurations . . . 31
  - 3.2.2 Classification Models . . . 31
  - 3.2.3 Evaluation Metrics . . . 32

3.2.4 Results . . . 32
3.3 Discussion . . . 34
3.4 Conclusion and Future Work . . . 35

4 AlgVec Word Embeddings . . . 36
4.1 Introduction . . . 36
4.2 Research Objectives . . . 37
4.3 Building the AlgVec Embeddings . . . 38
4.3.1 Data Collection . . . 38
4.3.2 Data Preprocessing . . . 38
4.3.3 Embedding Techniques Used in AlgVec . . . 39
4.4 Experimental Evaluation and Results . . . 42
4.4.1 Experimental Environment . . . 42
4.4.2 Intrinsic Evaluation . . . 44
4.4.3 Sentiment Analysis Task . . . 48
4.5 Discussion . . . 53
4.5.1 Intrinsic tasks . . . 53
4.5.2 Extrinsic tasks . . . 54
4.6 Conclusion . . . 56

5 Arabizi Sentiment Transfer Learning . . . 59
5.1 Introduction . . . 59
5.2 Proposed Architecture . . . 60
5.3 Experiments . . . 61
5.3.1 Dataset Preparation . . . 61
5.3.2 Fine-Tuning for Sentiment Analysis . . . 61
5.3.3 Evaluation Metrics . . . 62
5.3.4 Baseline Comparisons . . . 62
5.4 Results and Discussion . . . 62
5.5 Conclusion . . . 63

6 Algerian Dialect Sentiment Benchmark . . . 66
6.1 Introduction . . . 66
6.2 Methodology . . . 67
6.2.1 Data Collection and Annotation . . . 67
6.2.2 Data Preprocessing . . . 68
6.2.3 Data Representation . . . 72
6.2.4 Sentiment Classification Models . . . 74
6.3 Experiments and Results . . . 76
6.3.1 Model Evaluation . . . 76
6.3.2 First Experiment: Evaluating Preprocessing and Feature Extraction . . . 77
6.3.3 Second Experiment: Comparing Word Embedding Models for DL Architectures . . . 79
6.3.4 Third Experiment: MARBERT with LSTM Hybrid Model . . . 80
6.3.5 Cross-Domain Evaluation . . . 81
6.4 Discussion . . . 82
6.4.1 Comparison with prior studies . . . 82
6.4.2 Cross-Domain Performance Discussion . . . 83
6.4.3 Error Analysis and Model Limitations . . . 84

6.5 Conclusion . . . 85

7 Integrating Sentiment Analysis into HMI . . . 87
7.1 Introduction . . . 87
7.2 The Role of Sentiment as an Interaction Signal . . . 88
7.3 System Architecture and Integration . . . 88
7.3.1 Linguistic Preprocessing for Dialectal Robustness . . . 88
7.3.2 Sentiment Modeling . . . 89
7.4 Decision Layer for Adaptive Interaction . . . 89
7.4.1 Sentiment to Action Mapping . . . 89
7.4.2 Contextual Response Adaptation . . . 89
7.5 Conclusion . . . 90

8 Conclusion and Future Works . . . 91

A Annotation Guidelines . . . 107
A.0.1 Annotator Disagreements . . . 107
A.0.2 Inter-Annotator Agreement . . . 107

B Arabizi to Arabic Mapping . . . 109

List of Figures

2.1 Multifaceted approach to sentiment analysis. Different techniques are applied across various text context levels, illustrating how each level enhances sentiment understanding and rating granularity. 9
2.2 Aspect level 10
2.3 Sentiment Analysis taxonomy 13
2.4 The role of X in sentiment analysis. The X platform provides real-time access to vast user-generated data, supporting applications like trend analysis, brand monitoring, and public opinion research. 14
2.5 Arabic language varieties 15
2.6 Hierarchy of Algerian dialect varieties 16
2.7 Challenges of sentiment analysis for Arabic and its dialects on X. The figure highlights three key difficulty areas distinguished by color: X platform-specific challenges (blue), sentiment analysis limitations (orange), and Arabic language complexities (yellow). 18

3.1 Classification performance comparison (Accuracy and Macro F1-score) of SVM and LSTM models with different stopword removal methods. 35

4.1 Overview of the AlgVec development pipeline, including data preprocessing, embedding training and evalutaion tasks. 43
4.2 t-SNE visualization of sentiment word embeddings showing clustering of sentiment and dialect-specific terms in Algerian dialect. 49
4.3 Performance comparison of embedding models (AlgVec-CBOW, AlgVec-Skip-gram, AlgVec-FastText, AraVec, MARBERT with AlgVec variants) on SA using SVM, CNN, and LSTM classifiers. 55

5.1 Overview of the proposed hybrid framework for Arabizi sentiment analysis . . . 61

6.1 Important events that have occurred in Algeria between 2018 and 2023 68
6.2 Flow diagram of the proposed architecture, illustrating the main components and data flow for Algerian sentiment analysis on X. 75
6.3 Impact of Preprocessing Techniques on the Performance of SVM and MNB Models 78
6.4 Performance of MARBERT-LSTM on Cross-Domain Sentiment Classification and Improvement after SMOTE Oversampling 84

7.1 The proposed sentiment-aware HMI prototype for Algerian dialect input. The system displays the predicted sentiment and adaptive response in real time. . . 90

List of Tables

2.1 A Comparison of Different Arabic Dialects for the English expression How are you today? . . . 16
2.2 Summary of recent studies on Word Embeddings for Arabic and its dialects . . 22
2.3 Summary of Related Work on Sentiment Analysis for Arabic and Arabic Dialects 27
2.4 Overview of Sentiment Analysis Datasets for Arabic and Arabic Dialects 28
2.5 Overview of Sentiment Analysis Datasets for Algerian Dialect 28

3.1 Examples of Algerian Arabic Stopwords with Explanations 31
3.2 SVM Classification Report with Different Stopword Removal Approaches 33
3.3 LSTM Classification Report with Different Stopword Removal Approaches . . . 34

4.1 Statistical summary of the collected Algerian dialect X corpus. 38
4.2 Example word pair similarity scores in Algerian dialect using AlgVec embeddings. 46
4.3 Comparative cosine similarity scores across semantic relations in Algerian dialect embeddings (Skip-gram, CBOW, FastText). 47
4.4 Top-5 nearest neighbors of sentiment words in Algerian dialect embeddings. . . 48
4.5 LSTM and CNN model hyperparameters. 51
4.6 SVM classification results using different word embeddings on the Algerian dialect dataset. 51
4.7 CNN and LSTM classification results using different embeddings on the Algerian dialect dataset. 52
4.8 Classification results using MARBERT-only embeddings on the Algerian dialect dataset. 52
4.9 SVM classification results using MARBERT + AlgVec embeddings on the Algerian dialect dataset. 53
4.10 CNN and LSTM results using MARBERT + AlgVec embeddings on the Algerian dialect dataset. 53
4.11 Examples of Misclassified Tweets 58

5.1 Summary of Experimental Results on NArabizi Dataset 64
5.2 Ablation study illustrating the contribution of each component to overall performance (measured by Macro F1). 65

6.1 Annotated Examples from the Algerian Dialect Dataset Illustrating Linguistic Challenges and Sentiment Ambiguities 69
6.2 Examples of tweets before and after the preprocessing pipeline 73
6.3 Baseline results for SVM and MBN with different feature configurations 77
6.4 Performance comparison between our best model and the best model from Albalawi et al. (2021) 79
6.5 BiLSTM and CNN Model Hyperparameters 79
6.6 Comparison of CNN and BiLSTM models across different embeddings techniques 80

6.7 Comparative Performance Metrics of all models 81
6.8 Performance of MARBERT-LSTM on External Datasets, Highlighting SMOTE Oversampling Effects 82
6.9 Comparative performance analysis of the proposed model and previous studies . 83
6.10 Examples of Misclassified Tweets for Error Analysis 85

7.1 Mapping sentiment predictions to interaction strategies. 89

A.1 Sample Tweets illustrating Annotation Disagreements for Algerian Sentiment Analysis 108

B.1 Sample Arabizi-to-Arabic mapping entries used for Normalization 109

Nomenclature

ADSA	Algerian Dialect Sentiment Analysis
ASA	Arabic Sentiment Analysis
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DL	Deep Learning
HMI	Human-Machine Interaction
LSTM	Long Short-Term Memory
ML	Machine Learning
MSA	Modern Standard Arabic
NB	Naïve Bayes
NLP	Natural Language Processing
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency

Chapter 1

Introduction

HMI has become a fundamental pillar of modern technological advancements, closing the gap between humans and intelligent systems. The ability of machines to understand and interpret human sentiments and subjective expressions is crucial for creating systems that are intuitive, adaptive, and user-centric. Sentiment analysis, a subfield of NLP, plays a central role in enabling machines to comprehend and respond to human emotional states effectively. Despite significant progress, challenges posed by language diversity, particularly in dialectal variations, remain largely underexplored.

1.1 Context and motivations

X (formerly Twitter)[1] has become a major platform for individuals to express opinions, emotions, and sentiments on a wide range of topics (Hajirahimova and Ismayilova, 2022). As a result, large volumes of subjective and unstructured textual data are generated daily (Wang et al., 2022). Analyzing such data provides valuable insights for various applications, including marketing, political analysis, social media monitoring, and brand reputation management. Owing to its real-time nature, X is particularly suitable for tracking public sentiment and emerging social trends (Zimbra et al., 2018). Consequently, significant research efforts have focused on improving sentiment analysis techniques for X data (Wang et al., 2022).

Sentiment analysis is one of the most active research areas in NLP, aiming to automatically identify and classify subjective information such as emotions, opinions, attitudes, and evaluations. This task has been addressed using both classical ML approaches (Hajirahimova and Ismayilova, 2022) and more recent DL-based models (Yadav and Vishwakarma, 2020). Beyond content analysis, sentiment analysis plays a crucial role in enabling systems to understand users at an affective level.

In this context, HMI represents a key application domain for sentiment analysis. By

[1] https://twitter.com/home

allowing machines to infer users' emotional states from textual input, sentiment-aware systems can adapt their responses and behavior accordingly, leading to more natural, intuitive, and user-centered interactions. Such capabilities are essential for conversational agents, recommendation systems, and interactive platforms that rely on continuous communication between humans and machines.

Recent advances in affective computing further emphasize the importance of incorporating emotional and cognitive awareness into HMI systems. Beyond task-oriented processing, intelligent systems are increasingly expected to adapt to users' emotional states and mental workload in real time in order to improve interaction quality, usability, and user satisfaction. As highlighted by (Iarlori et al., 2024), the integration of affective state recognition with cognitive workload assessment represents a key direction for developing more natural, empathetic, and user-centered HMI systems. This perspective reinforces the role of sentiment analysis as a fundamental component for enabling emotionally aware AI systems capable of adapting their behavior based on users' subjective states.

Arabic is among the most widely studied languages in sentiment analysis, ranking after English and Chinese (Yadav and Vishwakarma, 2020), due to its increasing presence on social media platforms. Arabic is the eighth most used language on X, accounting for approximately 1.2% of all public tweets, corresponding to nearly 2.2 million tweets per day[2]. Traditionally, most Arabic NLP resources and sentiment analysis tools have focused on MSA (Oueslati et al., 2020). However, real-world user interactions on social media predominantly involve dialectal Arabic, which remains under-resourced and linguistically complex (Shamsi and Abdallah, 2021).

Among these dialects, Algerian Arabic (Darija) is widely spoken in Algeria and increasingly used on X, with approximately 1.15 million users as of early 2023[3]. Algerian Darija is characterized by non-standard orthography, frequent code-switching, and informal expressions (Yadav and Vishwakarma, 2020). Due to the lack of annotated datasets and specialized linguistic resources, ADSA remains a challenging and relatively underexplored research area (Oueslati et al., 2020). These challenges directly impact the effectiveness of sentiment-aware HMI systems, highlighting the need for dedicated models, representations, and preprocessing techniques tailored to Algerian dialect text.

Motivated by these challenges, this thesis aims to enhance human-machine interaction by improving sentiment analysis for the Algerian dialect, thereby enabling machines to better understand users' emotions and respond in a more adaptive and user-centered manner.

[2]https://semiocast.com/arabic-highest-growth-on-twitter
[3]https://datareportal.com/reports/digital-2023-algeria

1.2 Research Questions and Research Aims

1.2.1 Research questions

To enhance the understanding of ADSA on X data, the following research questions (RQs) were formulated:

- Why is X an appropriate platform for studying ASA?
- What are the main challenges faced by ASA and, more specifically, Algerian Dialect Sentiment Analysis?
- What publicly available datasets exist for ADSA, and how can a large open corpus be constructed for this low-resource language?
- What is the most effective approach for handling Arabizi in sentiment analysis tasks?
- How do different preprocessing techniques applied to Arabic tweets influence ASA and ADSA performance?
- What is the impact of removing or replacing emojis on ADSA accuracy?
- Which word embedding models and combinations yield the best performance when used with deep learning architectures?
- What machine learning and deep learning techniques have been applied to ASA and ADSA, and how do they compare in terms of performance?
- How can sentiment analysis be integrated into human-machine interaction systems to support adaptive and emotion-aware interaction for Arabic-speaking users?
- What challenges arise when deploying ADSA models in real-time interactive systems, and how can these challenges be addressed?

1.2.2 Research Aims

The main aim of this thesis is to develop and evaluate sentiment analysis approaches for Algerian Arabic dialect tweets, with the objective of enhancing sentiment-aware human-machine interaction. To achieve this aim, the following objectives are pursued:

- Construct and annotate a dataset of Algerian Arabic dialect tweets for sentiment classification into positive, negative, and neutral categories.
- Apply and analyze preprocessing techniques to effectively handle dialectal variations, informal language, and code-switching.

- Implement and benchmark a range of sentiment analysis models, including classical machine learning methods and state-of-the-art deep learning approaches.
- Investigate dialect-specific linguistic features and develop large-scale word embeddings for Algerian Arabic and Arabizi trained on extensive tweet collections.
- Design and implement a sentiment-aware interaction prototype that integrates the proposed sentiment analysis models within a human-machine interaction framework.
- Analyze the contribution of sentiment-aware mechanisms to interaction adaptivity and system responsiveness through qualitative analysis and performance-based evaluation.

1.3 Research Contributions

This thesis makes several contributions to the fields of sentiment analysis, Arabic NLP, and human-machine interaction, with a particular focus on the Algerian dialect.

Technical Contributions

- The construction and annotation of a sentiment-labeled dataset of Algerian Arabic dialect tweets, addressing the scarcity of resources for this low-resource language variety.
- The development and evaluation of preprocessing techniques tailored to Algerian Arabic social media content, including handling dialectal variations, informal expressions, and code-switching.
- An extensive experimental study of sentiment analysis models, including classical machine learning methods and deep learning architectures, with a particular focus on transformer-based approaches combined with dialect-specific word embeddings.

Representation and Modeling Contributions

- The introduction of **AlgVec**, a set of large-scale word embedding models for Algerian Arabic and Arabizi, trained on millions of tweets and evaluated through intrinsic and extrinsic sentiment analysis tasks.
- An in-depth analysis of the impact of word embedding choices and preprocessing strategies on sentiment classification performance for Algerian dialect text.

Human-Machine Interaction Contributions

- The design of a sentiment-aware interaction framework that integrates Algerian dialect sentiment analysis into human-machine interaction systems.

- The implementation of a graphical prototype demonstrating how sentiment predictions can be exploited to adapt system responses, illustrating the role of affective information in improving interaction responsiveness.

Overall, this research demonstrates how advances in sentiment and subjectivity analysis for Algerian Arabic can be leveraged to enhance human-machine interaction. By enabling machines to interpret users' emotional states expressed in dialectal language, the proposed framework contributes toward more adaptive, emotion-aware, and linguistically inclusive interactive systems. The findings help bridge the gap in sentiment analysis for low-resource Arabic dialects and provide practical insights for integrating affective intelligence into real-world applications such as conversational agents and interactive support systems.

1.4 Thesis plan

In addition to the current chapter, this thesis is organized as follows:

- Chapter 2 provides a thorough overview of sentiment and subjectivity analysis, covering the different levels and techniques. It also delves into the specifics of the Arabic language, dialects, and the challenges in sentiment analysis for Arabic, especially the Algerian dialect. The discussion of human-machine interaction and the review of relevant literature are also relevant to the thesis topic.

- Chapter 3 presents an attention-based approach for stopword extraction in Algerian Arabic. It covers the methodology for identifying influential words, the construction of custom stopword lists, for Algerian dialect and the impact of stopword removal on text classification performance.

- Chapter 4 introduces **AlgVec**, a set of word embedding models for Algerian Arabic and Arabizi. This chapter details the data collection, preprocessing, embedding training techniques (Word2Vec and FastText), and intrinsic evaluation of the embeddings, including nearest-neighbor analysis and word similarity tasks.

- Chapter 5 focuses on sentiment analysis for Arabizi text using transfer learning. It presents the model architectures, fine-tuning strategies, dataset preparation, and evaluation metrics for predicting sentiment in Arabizi written in the Algerian dialect.

- Chapter 6 concentrates on the core experiments combining **AlgVec** embeddings and transformer-based models (e.g., MARBERT + LSTM) for sentiment classification. It compares classical machine learning approaches and deep learning architectures, analyzes the results, and discusses cross-domain performance and error analysis.

- Chapter 7 focuses on integrating sentiment analysis into human-machine interaction. It presents a sentiment-aware framework that adapts system behavior based on emotions inferred from Algerian dialect text. The chapter outlines the interaction architecture and introduces a graphical prototype as a proof-of-concept to demonstrate how sentiment-aware mechanisms can enhance interaction quality and user experience.

- Chapter 8 summarizes the key findings and contributions of the thesis. It highlights the impact of the proposed methods on human-machine interaction and sentiment analysis in Algerian Arabic, and outlines potential directions for future research.

1.5 Publications

The following papers are part of the research carried out during the Ph.D. program:

1.5.1 International journals

1. Ouchene, L., Bessou, S. Benchmarking sentiment analysis of algerian arabic dialect on X (Twitter). J. King Saud Univ. Comput. Inf. Sci. 37, 358 (2025). `https://doi.org/10.1007/s44443-025-00295-w`.

2. Ouchene, L., Bessou, S. AlgVec: A word embedding model for the algerian dialect in arabic and arabizi. J. King Saud Univ. Comput. Inf. Sci. (2025). `https://doi.org/10.1007/s44443-025-00407-6`.

1.5.2 International conferences and workshops

1. L. Ouchene and S. Bessou, "Sentiment Analysis for Algerian Dialect Tweets," *14th International Conference on Computational Intelligence and Communication Networks* (CICN), Al-Khobar, Saudi Arabia, 2022, pp. 235-239, `https://doi.org/10.1109/CICN56167.2022.10008314`.

2. L. Ouchene and S. Bessou, "FastText Embedding and LSTM for Sentiment Analysis: An Empirical Study on Algerian Tweets," *International Conference on Information Technology* (ICIT), Amman, Jordan, 2023, pp. 51-55, `https://doi.org/10.1109/ICIT58056.2023.10226060`.

3. L. Ouchene and S. Bessou, "Improving Dialectal Text Classification via Attention-Based Stopword Extraction for Algerian Arabic," *2nd International Conference on Artificial Intelligence and Its Applications* (ICAITA), Mascara, Algeria, Oct 20th, 2025.

4. L. Ouchene and S. Bessou, "Self-Supervised Transfer Learning for Arabizi Sentiment Analysis," *2nd International Workshop on Machine Learning and Deep Learning* (WMLDL25), Msila, Algeria, Nov 25th, 2025. (Book of Abstracts).

5. L. Ouchene and S. Bessou, "X for Algerian Dialect Sentiment Analysis: A New Dataset Resource," *Scientific Research Day on Recent Advances in Arabic Language Processing* (RAANLP), University of Setif 1, Setif, Algeria, May 15th, 2025.

Chapter 2

Background

2.1 Introduction

In this chapter, we provide a brief overview of the Arabic language from both linguistic and computing perspectives. We discuss the challenges and specific characteristics of the Arabic language that make it unique in NLP research and how these differ from those found in other languages. Our goal is to provide a comprehensive understanding of the complexities and nuances of the Arabic language and its impact on NLP research and development. After that, we review the state-of-the-art deep learning models and different word representation techniques that have been developed for NLP tasks. We provide an explanation of these models and techniques, highlighting their strengths and limitations. Finally, we mention related work that have inspired our publications and contributed to the advancement of NLP research in the Arabic language. The techniques discussed here are a selected subset relevant to our research application. There are many additional techniques that have been developed in the field of Arabic NLP that could potentially be useful for various tasks and applications.

2.2 Sentiment Analysis: An Overview

Sentiment analysis is an NLP task focused on identifying subjective information in texts. While traditionally limited to polarity classification, SA has evolved to include emotion detection, subjectivity analysis, and fine-grained sentiment scoring (Cui et al., 2023). SA covers a range of aspects, taking into account factors such as techniques, textual context, and rating levels, as illustrated in Figure 2.1. Based on granularity, SA is typically performed at three levels:

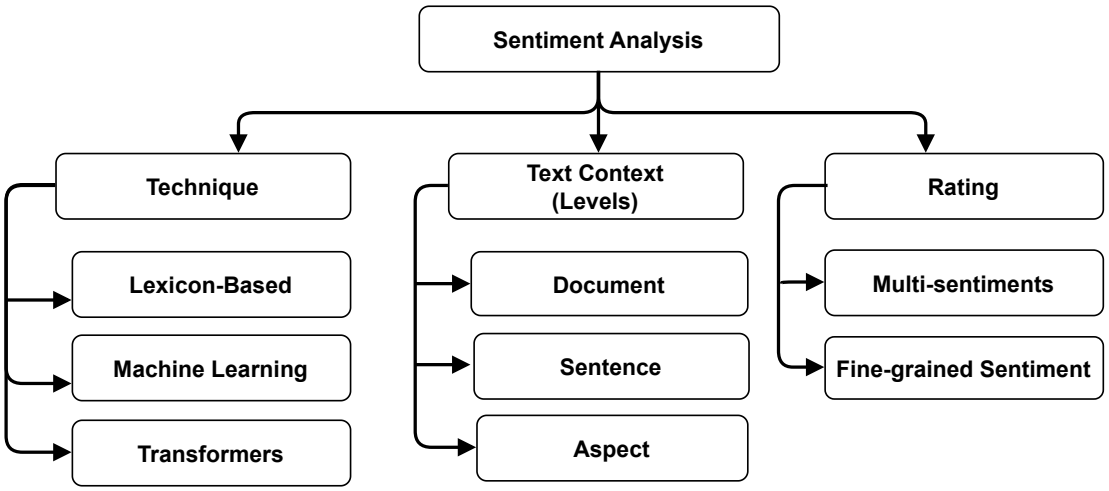


Fig. 2.1 Multifaceted approach to sentiment analysis. Different techniques are applied across various text context levels, illustrating how each level enhances sentiment understanding and rating granularity.

2.2.1 Levels of Sentiment Analysis

Currently, it is possible to apply a SA model at different levels of granularity based on specific requirements and scope. Typically, SA has been divided into Document Level, Sentence Level, and aspect Level.

Document level This level of SA aims to classify the overall sentiment expressed in an entire document as positive, negative or neutral (Behdenna et al., 2016). This variant of SA is less commonly utilized, it can be used to classify chapters or pages within a book (Mayur Wankhade and Kulkarni, 2022). Both supervised and unsupervised approaches can be used to classify the document (Bhatia et al., 2015). This field of SA poses a significant challenge due the presence of multiple contradictory polarities at the phrase and word levels. Existing approaches often overlook the local context or fail to consider both local and global context. In recent research, (Truică et al., 2021) proposed a novel approach called DocTopic2Vec to address these limitations. The model is evaluated against multiple Machine and Deep Learning models, showcasing improved performance compared to the baseline approach.

Sentence Level In Sentence-Level Sentiment Analysis (SLSA), each sentence is treated as an independent unit, and sentiment labels are assigned based on the subjective instances within each sentence (Feldman, 2013). The primary challenge at the sentence level is accurately interpreting the nuanced sentiment expressed, where context and negation can significantly impact the overall sentiment polarity. Negative words like 'not' have the potential to reverse or modify the sentiment associated with specific words, leading to a more nuanced sentiment analysis at the sentence level. For example, consider the

sentence: "The hotel room was clean, but the service was not speedy.". In this instance, the inclusion of the word 'not' alters the overall sentiment. Recent work in SLSA have achieved state-of-the-art performance through the utilization of deep learning models. Twitter typically employs SLSA to analyze user's tweets, aiming to extract insights into social trends and public opinion. Various deep neutral network DNN models have successfully achieved state-of-the-art performance in SLSA. In particular, the pre-trained models such as BERT, through semantic learning on large-scale corpus, can automatically implicit sentiment features (Su et al., 2023).

Aspect Level Aspect-level, also known as feature-level analysis, aims to examine the sentiment expressed towards specific aspects or features within a text or sentence (Zhang et al., 2022). For example, consider the following sentence: 'Dr. X provides excellent patient care. The staff was kind, and I felt very comfortable'. In aspect-level analysis, we first identify and extract aspect terms, such as 'patient care', 'staff professionalism', and 'comfort'. Next, we identify and extract opinion terms like 'excellent,' 'kind,' and 'very comfortable.' Sentiment polarities (positive, negative, or neutral) are then assigned to these opinion terms based on their sentiment (see Figure 2.2). Finally, the sentiment polarities of different opinion terms are aggregated to determine the overall sentiment for the entire text or specific aspects. Recent advancement in DL have enhanced the popularity of aspect-level by enabling a more quantifiable analysis of semantic relationships between words and the aspects (specific features) they describe (Cui et al., 2023).

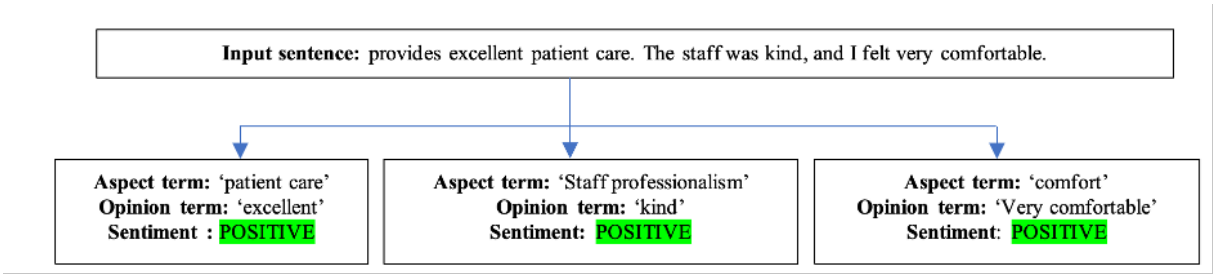


Fig. 2.2 Aspect level

2.2.2 Techniques of Sentiment analysis

Approaches of SA refer to the techniques used to identify the sentiment. Current prevalent SA methods centered on sentiment dictionaries, sentiment rules, statistics-based machine learning models, neural network-based deep learning models, and pre-training models which are regrouped by (Cambria, 2016) into three primary groups: Knowledge-based, statistical, and hybrid methods. Knowledge-based approaches utilize sentiment lexicons for text classification, while statistical methods employ machine learning and deep learning models trained on annotated data. Hybrid methods combine the strengths of both.

Researchers are constantly working to improve the performance of SA methods, aiming for reduced computational cost and higher accuracy. The selection of SA model depends on various factors, such as the domain and purpose of the task, the availability of computational resources, the quality and size of training data, and the accuracy requirements. Generally, it is advisable to test various models to determinate which one yields optimal performance for the specific task application. Figure 2.3 presents an improved visualization of SA taxonomy, inspired by (Mayur Wankhade and Kulkarni, 2022; Abdullatif Ghallab and Ali, 2020). In the following section, we discuss the most popular SA models for social media.

- **Lexicon-Based Approach** , also known as sentiment lexicon, determines the emotional polarity of text using predefined dictionaries of words and phrases with associated sentiment scores (Bonta et al., 2019). While easy to implement and requiring less training data than machine learning-based methods, it struggles with adaptability across different domains due to varying sentiment connotations of words in different contexts.

- **Machine Learning-Based Approach** is the most prevalent approach to SA, requiring labeled data for supervised learning and including traditional models like **SVM** , **NB**, and Deep Learning models such as **CNN**, **RNN)**, and **LSTM**, which are among the most popular models (Cui et al., 2023; Alexander Ligthart and Tekinerdogan, 2021). Traditional models require a time-consuming feature engineering process (Yadav and Vishwakarma, 2020). However, deep learning models automatically learn complex features and optimize their hyper-parameters, leading to improved accuracy and performance (Dang et al., 2020). These models offer superior handling of unstructured data from social media and capture complex patterns, enhancing overall accuracy.

- **Transfer learning** is an innovative technique in machine learning that leverages existing knowledge obtained from a large-scale source domain to address diverse challenges and achieve exceptional prediction accuracy (Liu et al., 2019). A popular technique in recent years within transfer learning is the Pretrained Language Model.

  Pre-trained Language Models (PMLs) use self-supervised learning on large datasets and fine-tuning to excel in various tasks (Wang et al., 2023), with mechanisms like those found in Transformers, which include self-attention and positional encoding, enabling a deeper understanding of text and achieving state-of-the-art results in sentiment analysis (Vaswani et al., 2023).

  The transformer introduces novel mechanisms, including self-attention and positional encoding, to understand relationships between words over long distances, leading to a deeper textual understanding. Self-attention computes “attention

weights" for every pair of word in a sequence, enabling the transformers to focus on relevant segments for the representation of the sequence (Vaswani et al., 2023).

BERT, a prominent PLM proposed by Google in 2018, utilizes bidirectional transformers and employs a masked language model (MLM) pre-training objective, where some tokens in the input are masked, and the model predicts the original vocabulary id of the masked word based on the context provided by the unmasked tokens (Devlin et al., 2019). BERT outperforms traditional machine learning algorithms, achieving state-of-the-art results in sentiment analysis on social media data by leveraging its capability to capture nuanced language and contextual meaning (Bikku et al., 2023).

- **Hybrid approach** combines sentiments lexicons with machine learning methods to create a more flexible and accurate sentiment analysis model. It may also involve integrating rule-based systems with deep learning models or using domain-specific knowledge to improve sentiment analysis accuracy in specific contexts.

- **Multimodal SA** The rise of social media has led to the development of multimodal sentiment analysis. Traditional text-based SA, which focuses on analyzing the sentiment expressed in written language, has been limited in capturing the full range of sentiment expressed by individuals. Multimodal SA goes beyond text and incorporates additional modalities, such as facial and vocal expressions, presenting promising potential for effectively analyzing social media data where people express their sentiments in diverse ways (Soleymani et al., 2017).

2.3 The Role of X-Platform in Sentiment Analysis

X is a widely used micro-blogging platform that enables real-time communication through 280-character messages. It remains a key source of user-generated content for SA, providing diverse linguistic data through its API v2. Features[1] such as hashtags and geolocation offer rich contextual metadata for research across various topics, as illustrated in Figure 2.4. However, the prevalence of slang, abbreviations, and evolving language on the platform presents challenges that require advanced models capable of interpreting short but contextually rich messages (Bouassida and Mezali, 2025).

2.4 Arabic Language and Arabic dialects

Arabic (**العربية Al-Arabiya)** is primarily spoken in the Arabic world, covering twenty-two countries across the Middle East and North Africa. It is the native language of over 370

[1]https://help.twitter.com/en/resources/glossary

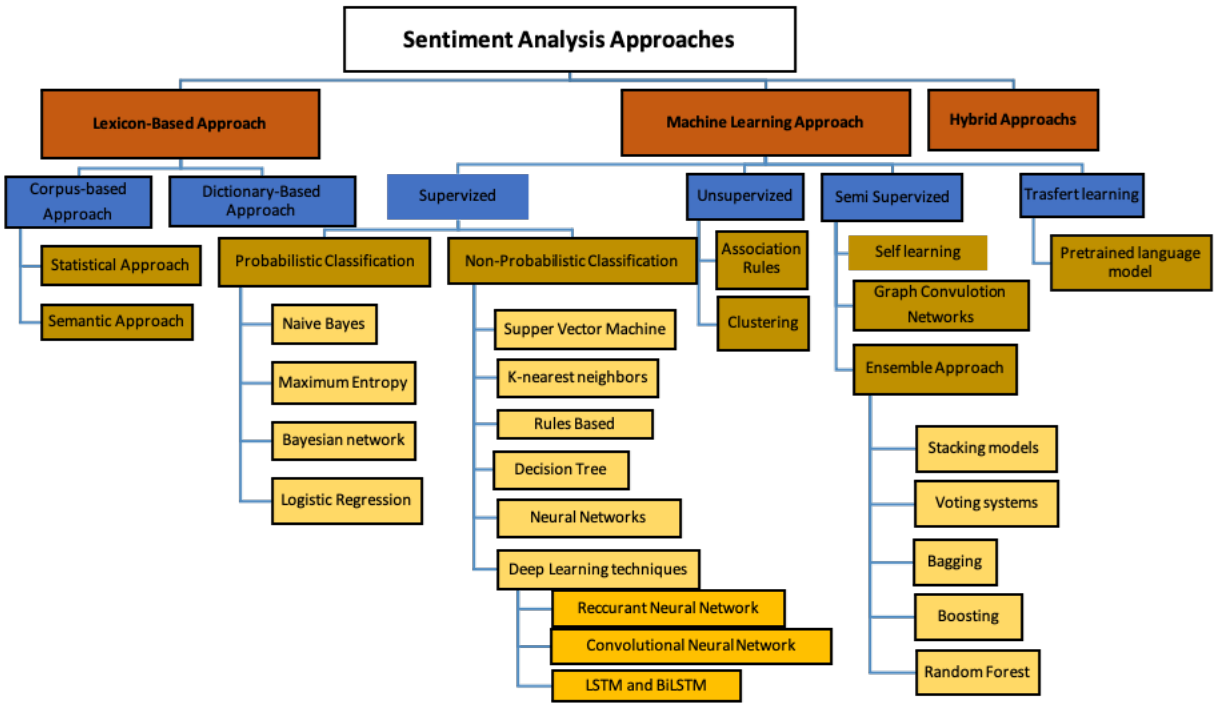


Fig. 2.3 Sentiment Analysis taxonomy

million [2] people and serves as one of the six official languages of the United Nations [3]. The Arabic alphabet consists of 28 letters, written from right to left, and lacks capital letters. Three main varieties of Arabic language coexist (See Figure 2.5), each serving a specific social purpose:

- **Classical Arabic (CA)** is closely associated with the Quran and finds application in many ancient Arabic manuscripts. It constitutes a highly formal language that is no longer used in everyday conversation. However, changes resulting from Arab interactions with non-Arab cultures, the demands of modernity, and the need to unite the Arabic word under a common language have led the emergence of new linguistic varieties alongside CA (Al-Huri, 2015).

- **MSA** (الفصحى Al-Fus'ha), a descendent of CA with subtle stylistic and lexical differences (Britannica, 2023) , has become the official language of Arab countries and is used for formal communication in government, media, education, and literature. However, native Arabic speakers speak different dialects, varying socially and linguistically from one area to another.

- **Dialectal Arabic (DA)** (لهجة La'hdja), referred to as colloquial (العامية Al-Aamyyah)

[2]https://www.worlddata.info/languages/arabic.php
[3]https://www.un.org/en/our-work/official-languages

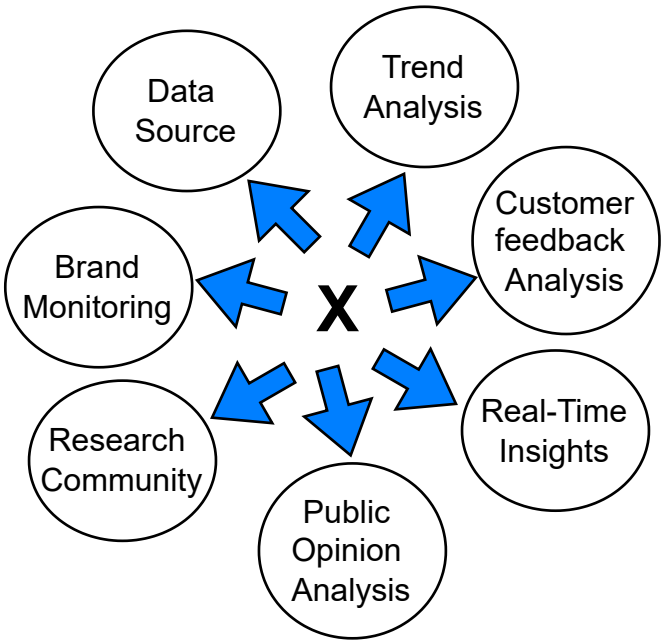


Fig. 2.4 The role of X in sentiment analysis. The X platform provides real-time access to vast user-generated data, supporting applications like trend analysis, brand monitoring, and public opinion research.

or (الدارجة Al-Da'rijah) , functions as a low and informal variety, acquired naturally with no need to learn grammar, used for people's day-to-day communication. In comparison to MSA, Arabic dialects have a simpler grammatical structure and significant variations in pronunciation, as shown in Table 2.1. This results in linguistic diversity within individual countries and across the Arab world. With over 30 Arabic dialects [4] organized into five groups, some are associated with specific countries, such as Egyptian and Iraqi, while others are linked to regions. For instance, Gulf Arabic is prevalent along the Gulf Coast, from Kuwait to Oman and Yemen. Additionally, Maghrebi Arabic is spoken in Algeria, Tunisia, Morocco, Mauritania and western Libya. Furthermore, Levantine Arabic is commonly heard in Palestine, Lebanon, Syria, and western Jordan. Moreover, Saudi Arabia features regional dialects such as Hijazi Arabic in the Hijaz region and Najdi Arabic in the Najd region's desert and oasis areas (Alnosairee and Sartini, 2021).

[4]https://iso639-3.sil.org/code/ara

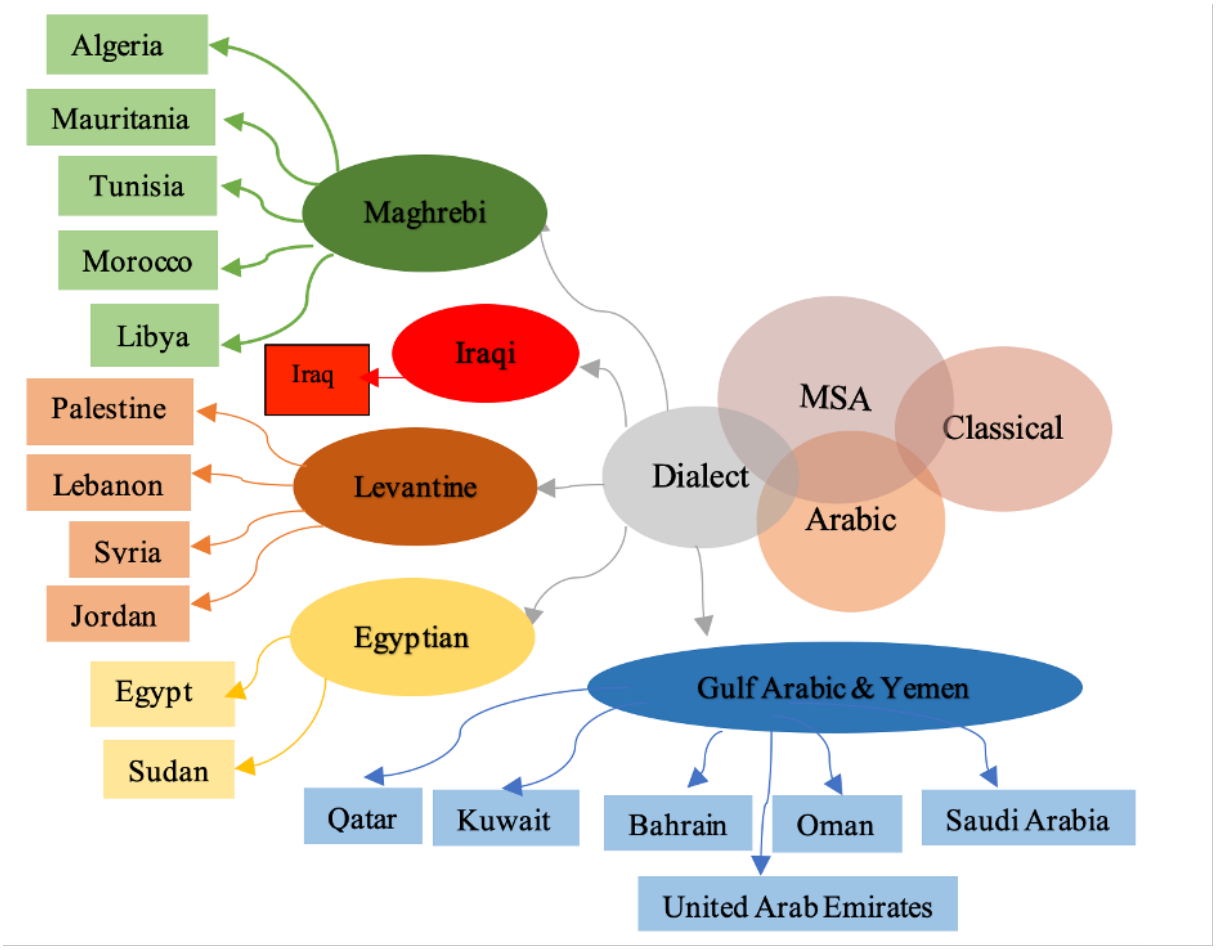


Fig. 2.5 Arabic language varieties

2.5 Algerian Dialect

The Algerian dialect, commonly known as ad-Dārija al-Jazairia, is spoken by over 46 million people[5] in Algeria. It comprises two main linguistic groups: Tamazight (Berber) and Algerian Arabic (the dominant spoken language), which evolved through Arabization and historical influences (Bougrine et al., 2017). Algerian Arabic has been significantly influenced by other languages, particularly French, which remains widely used due to colonial history (Rahmoun-Mrabet, 2018). This influence has led to linguistic characteristics such as code-switching. Additionally, Algerian Arabic exhibits notable regional variations, with distinct dialects spoken in the west (Mezzoudj et al., 2019), east, central (Harrat et al., 2016), and southern regions, as shown in Figure 2.6. These variations reflect the country's rich sociolinguistic diversity (Benrabah, 2013).

[5]https://www.ons.dz/

MSA	Dialectal Arabic	Group/Country
كيف حالك اليوم؟	واش راك اليوم؟ wāš rak lyūm ?	Maghrebi /Algeria
كيف حالك اليوم؟	إزيك النهاردة؟ ’izzīyyak in-nahār-dā?	Egyptian / Egypt
كيف حالك اليوم؟	كيفك اليوم؟ kyfīk lyūm?	Levantine/Syria
كيف حالك اليوم؟	شلونك اليوم؟ šlūnak lyūm?	Gulf / Saudi

Table 2.1 A Comparison of Different Arabic Dialects for the English expression How are you today?

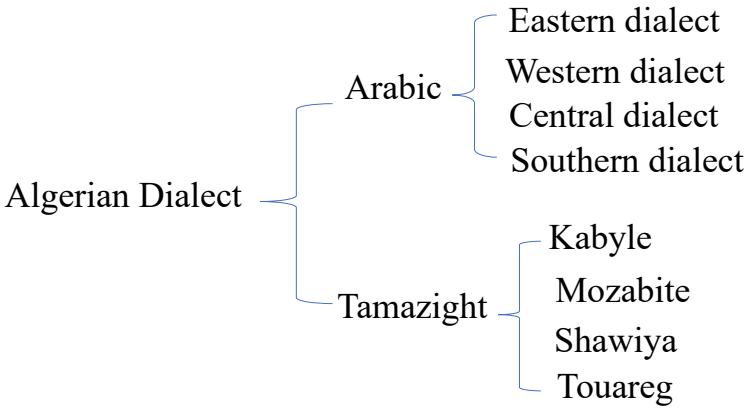


Fig. 2.6 Hierarchy of Algerian dialect varieties

2.6 Algerian and Arabic Sentiment Analysis Challenges on Social Media

The development of ADSA is influenced by three major factors, as illustrated in Figure 2.7.

2.6.1 Language Challenges

Algerian Arabic presents significant linguistic challenges due to its diverse regional dialects and lack of standardized spelling conventions. Several phenomena make automatic processing particularly difficult:

- **Code-switching** is common on social media, involving the mixing of French, English, and Arabic, even within a single comment (Shi and Agrawal, 2025). This linguistic blend complicates sentiment analysis by altering grammar, vocabulary, and sentiment cues (Dahou et al., 2025). For example, the phrase ”Les hommes ربي يوفقك” (”You are the man, may God help you succeed”) combines French and Arabic seamlessly. Our approach addresses these challenges through language-specific processing strategies, as described in Section 6.2.2.

- **Slang and Sarcasm** The Algerian dialect is rich in informal expressions and slang that vary by context, making interpretation difficult without cultural knowledge, especially for SA. For example, the sarcastic tweet "نتا بنادم ولا كيلو زيت مقلي؟" ("Are you a human or a liter of fried oil?") humorously mocks someone's appearance or behavior. Although playful, such expressions often carry implicit negative sentiment that may be overlooked by models not trained on dialect-specific language and social cues (Kumar et al., 2025).

- **Arabizi** is an informal writing system widely used on social media that represents Arabic words phonetically using Latin letters and numerals. For example, the numerals 3, 7, and 9 commonly stand for the Arabic letters Ain (ع), Ha (ح), and Qaf (ق), respectively. Despite its widespread use, research on Algerian Arabizi is limited (Abdedaiem et al., 2024a), mainly due to its lack of standardization. This high variability poses major challenges for text normalization and automated processing (Abd-Elnabi et al., 2023).

2.6.2 X-Specific Challenges

X's unique characteristics, such as its 280-character limit, encourage the use of abbreviations, slang, emojis, and hashtags, all of which complicate SA (Bouassida and Mezali, 2025). Furthermore, the platform's dynamic nature intensifies these challenges during real-time events like elections or crises (Ebrahimi et al., 2017).

2.6.3 Sentiment Analysis Challenges

Sentiment analysis faces inherent difficulties, such as detecting figurative language, especially sarcasm, which is common in both MSA and dialectal Arabic (Farha and Magdy, 2021). Furthermore, the absence of standardized tools and methods for preprocessing (Almutairi et al., 2024), combined with the scarcity of benchmark datasets (Oueslati et al., 2020), complicates the development of effective processing pipelines.

2.7 Human–Machine Interaction and Sentiment Interpretation

HMI is a multidisciplinary field that examines how humans interact with computers and intelligent systems to achieve effective, efficient, and satisfying communication. Early HMI research primarily concentrated on usability, accessibility, and interface design, focusing on how users issue commands and receive system feedback (Song, 2024). With the rapid advancement of artificial intelligence and NLP, however, the scope of HMI has expanded toward the development of intelligent interfaces capable of understanding, interpreting, and adapting to user behavior and affective states, moving beyond purely transactional interaction models (Carrasco Ramírez, 2024).

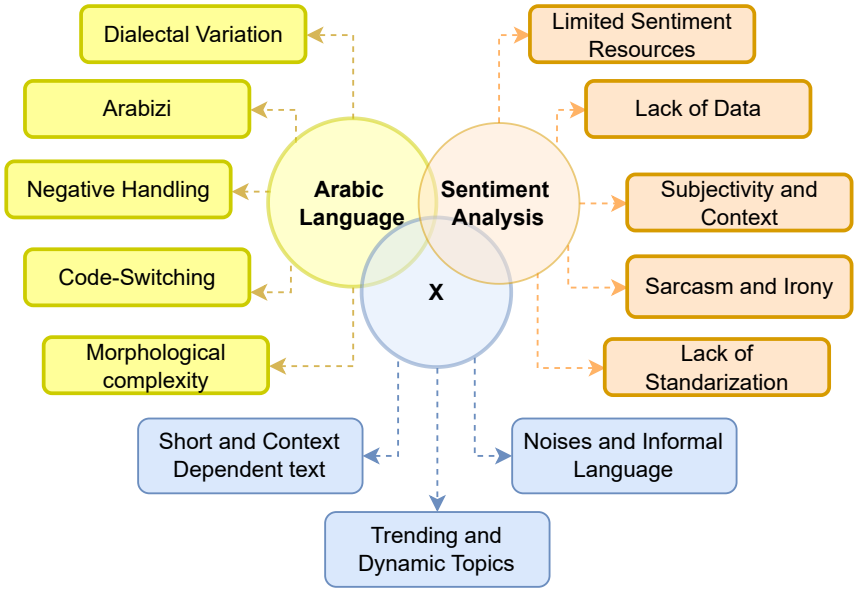


Fig. 2.7 Challenges of sentiment analysis for Arabic and its dialects on X. The figure highlights three key difficulty areas distinguished by color: X platform-specific challenges (blue), sentiment analysis limitations (orange), and Arabic language complexities (yellow).

Affective computing, a closely related research area, emphasizes the recognition and interpretation of human emotions to enhance human-centered systems. By enabling machines to perceive emotional cues in real time, affective computing contributes to more natural, responsive, and engaging interactions, ultimately improving user satisfaction. It is widely regarded as a foundational pillar of emotionally intelligent systems, where emotion recognition and sentiment analysis play a central role in interaction design (Carrasco Ramírez, 2024).

The ability to analyze users' subjective and emotional expressions has become increasingly important for advancing HMI systems. Research on sentiment-aware HMI suggests that incorporating emotional understanding into interaction mechanisms leads to more adaptive, responsive, and personalized systems. For instance, sentiment analysis has been widely applied in conversational agents, chatbots, and virtual assistants to infer users' emotional states, adjust system responses accordingly, and deliver more engaging and user-centered interactions (Ghosh, 2023; Hegde and Jayalath, 2025). This integration aligns with broader trends in AI research aimed at developing systems that understand not only the semantic content of language but also its emotional and contextual dimensions, thereby enhancing user trust and acceptance.

Within the field of affective computing, recent studies highlight the need for HMI systems that account for both emotional and cognitive factors during interaction. Rather than focusing solely on functional task execution, these systems aim to model users' affective states and mental workload to enable adaptive and human-centered behavior.

According to Iarlori et al. (2024), integrating affective state recognition with mental workload assessment provides a conceptual foundation for the development of more adaptive and emotionally intelligent interaction paradigms.

2.8 Related Work

The literature relevant to our contributions is reviewed in this section.

2.8.1 Related Work in Attention-Based Stopword Extraction

Stopword lists for MSA have been extensively studied, with several manually and statistically curated lists available (Sembok and Abuata, 2025). However, most Arabic dialect stopword lists rely on frequency-based heuristics or translations from MSA resources, which often fail to capture semantic irrelevance in context or adapt to task-specific needs (Alharbi, 2020).

For instance, (Medhat et al., 2015) generated an Egyptian Arabic stopword list from online social network data, demonstrating improved sentiment classification using Naïve Bayes and Decision Tree classifiers compared to MSA lists alone. Similarly, (El-Khair, 2017) compared general, corpus-based, and combined stopword lists for MSA, finding that combined lists performed best with BM25 weighting. A supervised method combining domain-independent and domain-specific corpora for context-aware stopword detection was proposed by (Namly et al., 2024).

Other efforts include (Alharbi and Lee, 2021), which released a multi-dialect stopword package based on the Twitter corpus Kawarith, and (NASSR et al., 2021), which developed a Moroccan dialect stopword list through a pipeline involving preprocessing, dictionary construction, and word embedding clustering on data from Facebook, Twitter, and YouTube.

Beyond Arabic, related work in other low-resource languages provides useful insights. For instance, (Bashir et al., 2024) constructed a Hausa stopword list using TF-IDF and manual validation on 1.2 million tokens from four news sources. The final list of 76 stopwords improved prior coverage by 6%, aiding tasks like sentiment analysis and machine translation. This highlights the value of combining statistical methods with expert review for effective stopword identification in under-resourced settings. More recently, (Kuwabara and Suzuki, 2024) proposed an attention-based method that identifies low-importance tokens across model predictions to dynamically generate stopword lists.

Despite these advances, no prior work has focused specifically on generating stopword lists tailored for the Algerian dialect.

2.8.2 Related Work in AlgVec Word Embeddings

The performance of deep learning and transformer-based models in NLP is strongly influenced by the quality of word representations. Prior work on Arabic embeddings can be grouped into three main themes: (1) general-purpose Arabic embeddings, (2) dialectal NLP resources, and (3) Arabizi-focused models.

General Arabic Embeddings

Early efforts in Arabic NLP adapted general embedding techniques such as Word2Vec and FastText to MSA. A key contribution is AraVec (Soliman et al., 2017a), a set of embeddings trained on over 3.3 billion words from Wikipedia, news, and X. AraVec explored CBOW and Skip-gram architectures, producing six models for downstream NLP tasks. (Abu Kwaik et al., 2019) used AraVec in a CNN-LSTM framework for SA, achieving 93.5% accuracy on 2,000 tweets, illustrating the benefits of domain-aware embeddings.

Subsequent studies compared embedding methods and architectures. (Elhassan et al., 2023b) showed that FastText (90%) and Word2Vec (86%) outperformed AraVec (79%) in SA, mainly due to AraVec's limited dialectal coverage. (Allahim and Cherif, 2024) improved results by training Word2Vec and FastText on curated Wikipedia and book corpora, reporting up to 99% sentiment accuracy and highlighting the impact of corpus quality and hyperparameter tuning. Transformer-based embeddings further advanced performance: AraBERT-LSTM (Alosaimi et al., 2024b) achieved over 97% accuracy on SA, including 90.4% on dialectal tasks, underscoring the strength of contextualized embeddings.

Dialectal NLP Resources

Addressing dialectal variation, resources such as DiaLex (Abdul-Mageed et al., 2021a) provide standardized benchmarks for multidialectal Arabic, showing that embeddings trained primarily on MSA often fail on dialect-specific semantic tasks. More recently, (Aboelela et al., 2025) evaluated a wide range of models, from classical machine learning (SVM, Naïve Bayes (NB), XGBoost) to deep learning (CNN, BiLSTM) and transformers (CAMeLBERT, XLM-RoBERTa, MARBERT). Their ensemble achieved state-of-the-art results with 90.4% accuracy, yet results also highlighted the persistent gap in dialect-aware embeddings.

A broader step toward dialect-specific modeling is Swan (Bhatia et al., 2025), a suite of Arabic-centric embedding models trained for ArabicMTEB, a large-scale benchmark for dialectal and cross-lingual NLP. Swan demonstrated that dialect- and culture-aware embeddings outperform general-purpose ones across semantic retrieval, classification, and bitext mining tasks.

Arabizi-Focused Models

Mazari and Djeffal (2022b) introduced a dataset of Algerian comments written in both Arabic script and Arabizi. They trained Word2Vec embeddings on this corpus, showing that a Multinomial Naïve Bayes classifier achieved 84.2% accuracy on Arabizi text, emphasizing the importance of dialect-specific and script-specific embeddings. Cross-lingual approaches have also been explored. (Lachraf et al., 2019) proposed ArbEngVec, a cross-lingual embedding model for Arabic-English trained on 93M parallel sentence pairs. While effective for word translation and bilingual NLP tasks, ArbEngVec does not directly support dialectal or Arabizi-specific text. DziriBERT (Abdaoui et al., 2022), a contextualized transformer model for Algerian Arabic, supports both Arabic script and Arabizi. However, it does not provide static embeddings for Arabizi, which remain useful for certain applications or as baselines. In summary, three trends emerge:

1. General Arabic embeddings achieve strong results but are primarily trained on MSA, limiting their effectiveness in dialectal contexts.

2. Dialectal resources such as DiaLex and Swan highlight the importance of dialect-aware embeddings, yet Algerian Arabic remains underrepresented.

3. Arabizi resources exist but are limited in dialect coverage or scope, and often lack static embeddings.

To date, no comprehensive embedding resource supports Algerian Arabic in both Arabic script and Arabizi. Table 2.3 summarizes recent studies and their key limitations.

2.8.3 Related Work in Arabizi Sentiment Analysis

Over the past decade, most NLP research on Arabic has focused on MSA and large spoken dialects such as Egyptian or Levantine, where resources and datasets are significantly more available (Shi and Agrawal, 2025). By contrast, work on Algerian Arabizi, with its unique linguistic influences and regional characteristics, remains limited and highly under-resourced (Touileb, 2022).

Early studies addressed Arabizi sentiment analysis with traditional machine learning methods. For example, (Chader et al., 2019) applied Naive Bayes and SVM classifiers to Algerian Arabizi, integrating preprocessing strategies such as vowel removal and transliteration to handle orthographic variation. Their experiments across three annotated datasets showed that SVM achieved the best results, improving the F-score from 76% to 87%. Similarly, (Hazimeh et al., 2024) proposed SAAD, a supervised framework for sentiment and abuse detection on Arabizi social media text, reporting 80% accuracy with a K-nearest neighbor model.

Table 2.2 Summary of recent studies on Word Embeddings for Arabic and its dialects

Study	Language	Embedding Model	Corpus	Evaluation Task	Limitations
AraVec (Soliman et al., 2017a)	MSA, Dialects	CBOW Skip-gram	3B	SA Named entity recognition	Weak on informal text; does not support Arabizi
MARBERT (Abdul-Mageed et al., 2021b)	MSA, Dialects	BERT-base	128M tweets	SA Named entity recognition Question Answering	Does not support Arabizi
ArbEngVec (Lachraf et al., 2019)	Arabic English	CBOW Skip-gram	93M parallel sentence pairs	Word Translation	Weak on informal text; does not support Arabizi
DiaLex (Abdul-Mageed et al., 2021a)	MSA, Dialects	CBOW Skip-gram FastText	3,070 word pairs	Word analogy tasks for syntactic	Does not support Arabizi
(Mazari and Djeffal, 2022b)	Algerian dialect	custom Word2Vec	466K comments	SA	Small corpus
(Allahim and Cherif, 2024)	MSA, Dialects	CBOW Skip-gram FastText	Watan Corpus (6M) Book Corpus (11M) Wiki Corpus (111M)	SA	Does not support dialectal social media or Arabizi text
Swan (Bhatia et al., 2025)	Arabic dialects	ARBERTv2 ArMistral	Unique assembly of datasets	Multi-task/dialect benchmark	Not Algerian-specific

Several recent initiatives have begun to fill the data gap for Arabizi. TUNIZI (Fourati et al., 2021) introduced the first large-scale Tunisian Arabizi sentiment dataset annotated for positive, negative, and neutral polarity, which has since become a benchmark for Tunisian Arabizi systems. FASSILA (Abdedaiem et al., 2024a) collected over 10K Algerian Arabic posts, including Arabizi, with labels for fake news and sentiment, offering a modern baseline suite. Another line of work focuses on leveraging social signals: (Bousri et al., 2022) explored rumor detection in Algerian Arabizi by combining rumor content with user interactions through attention mechanisms, showing that modeling social context improves performance. For Moroccan Darija, (Hajbi et al., 2024) developed a transliteration approach combining rule-based conversion with a weighted Levenshtein algorithm, achieving strong results across multiple datasets with 92.14% MRR and 88.44% accuracy.

To overcome the scarcity of semantic resources, Arabizivec (Mandalam et al., 2024), a word embedding model specifically designed for informal Arabizi sentiment analysis, outperformed baseline representations on multiple sentiment tasks. In parallel, (Tobaili et al., 2019) introduced SenZi, a lexicon for Lebanese Arabizi built by translating and transliterating existing lexicons and expanding them with word embeddings. SenZi contains 24.6k entries and was shown to improve sentiment classification on large Arabizi social media corpora.

More recently, researchers have moved toward Arabizi-native language models. BAERT (Baert et al., 2020), trained on millions of Egyptian Arabizi tweets, demonstrated significant improvements over multilingual BERT for sentiment analysis without relying on explicit transliteration. Similarly, DarijaBERT (Gaanoun et al., 2024) is a mono-dialect language model trained on approximately 100M Moroccan tokens across both Arabic script and Arabizi, with dedicated variants such as DarijaBERT-arabizi and DarijaBERT-mix. These models highlight the growing recognition of Arabizi as a distinct representation worthy of direct modeling.

2.8.4 Related Work in Arabic Sentiment Analysis

Evaluating SA performance across social media platforms is essential for understanding the effectiveness of various methods across languages and applications. Recent studies highlight the importance of SA on platforms like X for languages other than English, especially Arabic (Cui et al., 2023). To conduct a systematic literature review, we followed the structured search methodology outlined by (Kitchenham and Brereton, 2013). Our strategy involved targeted keyword searches on ASA and ADSA across multiple academic databases, employing the following search terms:

(”Arabic” OR ”Algerian Arabic Dialect” OR ”Arabizi”) AND (”Sentiment analysis”) AND (”Social Media” OR ”Twitter” OR ”X”) AND (”Machine Learning” OR ”Deep Learning” OR ”Supervised Learning” OR ”Transformers”) AND (”Word Embedding” OR

"Word Representations") AND ("Polarity" OR "Subjectivity" OR "Opinion Mining")

A preliminary screening based on titles and abstracts was conducted to eliminate irrelevant papers. After this step, 80 articles were selected for a full-text review.

Comparative Studies in Arabic Sentiment Analysis

Deep learning has significantly advanced ASA. For example, (Ombabi et al., 2020) achieved a 92.44% F1-score and 90.75% accuracy using a hybrid CNN-LSTM model with FastText embeddings across multiple domains. Similarly, (Mohammed and Kora, 2019) developed a corpus of 40,000 Arabic tweets, where LSTM outperformed CNN and RCNN, reaching 88.05% accuracy with data augmentation. A large-scale study by (Farha and Magdy, 2021) confirmed that deep learning models outperform traditional methods, particularly in handling sarcasm.

The rise of transformer-based models has further advanced ASA through self-supervised learning. Arabic-specific BERT variants like AraBERT (Antoun et al., 2021) and MARBERT (Abdul-Mageed et al., 2021b) have been key to this progress, notably achieving top results in the NADI 2022 competition (Abdul-Mageed et al., 2022). More recently, (Alosaimi et al., 2024a) introduced a model that combines AraBERT with LSTM to capture both contextual and sequential information. This hybrid approach outperformed traditional ML and DL methods, achieving 93.76% accuracy on four benchmark datasets.

Comparative Studies in Arabic Dialectal Sentiment Analysis

Interest in dialectal SA has grown, with studies addressing regional linguistic variations. (Elmasry et al., 2014) developed the Slang Sentimental Words and Idioms Lexicon (SSWIL), which, when combined with an SVM classifier, reached 87% accuracy on social media data. Similarly, (Kwaik et al., 2019) trained a BiLSTM-CNN model on multi-dialect Arabic sentiment, while (Farha and Magdy, 2019) introduced MAZAJEK, an online ASA system that reached 92% accuracy.

Several studies have targeted specific Arabic dialects. (Alomari et al., 2017) introduced the Arab Jordanian General Twitter (AJGT) corpus, achieving top results with SVM (88.72% accuracy and 88.27% F1-score). For the Sudanese dialect, (Abo et al., 2019) applied Decision Trees and Lexicon-based methods, reporting 83.5% and 60.85% accuracy, respectively. In Gulf dialects, (Alqarni and Rahman, 2023) studied 90,187 Saudi tweets on COVID-19, with CNN and BiLSTM models reaching 92.80% and 91.99% accuracy. Likewise, (Husain et al., 2022) created a Kuwaiti dialect dataset of 16.6k annotated tweets, where AraBERT surpassed traditional ML models, achieving 89% accuracy.

For Maghreb dialects, (Jbel et al., 2024) addressed the challenge of integrating Arabic and Latin scripts in Moroccan Arabic sentiment analysis, achieving 92% accuracy. Building on this regional focus, (Masmoudi et al., 2021) examined Tunisian Facebook

comments, applying CNN, LSTM, and Bi-LSTM models, enhancing the F1-score to 87% by incorporating morpho-syntactic features.

Complementing deep learning approaches, a comparative study by (Jaballi et al., 2022) showed that traditional ML methods such as NB, SVM, and LR, when combined with TF-IDF features, achieved over 84% accuracy. In the same study, Bi-LSTM models performed best, reaching 88% accuracy. Expanding beyond individual dialects, (Muhammad et al., 2023) introduced the Sentiment Analysis for African Languages benchmark. This resource covers 14 African languages, including Algerian and Moroccan Arabic, and evaluates both fine-tuned pre-trained models and classical ML techniques.

Comparative Studies in Algerian Dialect Sentiment Analysis

Algerian dialect sentiment analysis has progressed from early lexicon-based methods to more sophisticated deep learning and hybrid approaches. In the early stages, the lack of annotated datasets encouraged the use of lexicon-based techniques. For instance, (Mataoui et al., 2016) achieved 79.13% accuracy using a lexicon-based approach on a dataset of 7,698 Facebook comments.

Comparisons between classical ML and DL approaches have also been explored. (Soumeur et al., 2018) applied CNN and Multi-Layer Perceptron models to Algerian Facebook comments, achieving 89.5% and 81.6% accuracy, respectively. Similarly, (Moudjari et al., 2020) found that CNN and LSTM models reached 76% accuracy, while (Mazari and Djeffal, 2022a) reported that SVM outperformed CNN and RNN, suggesting that SVM is well-suited for binary classification on moderately sized datasets. Event-specific analyses have also appeared; for instance, (Mazari and Djeffal, 2021) analyzed 7,800 comments related to the 2019 Hirak protests. They obtained 63.28% accuracy with CNN and 60.97% with RNN. The work of Ouchene and Bessou (2022) focused on sentiment analysis of Algerian Arabic tweets, addressing challenges related to dialectal variation and informal language. In a subsequent study, Ouchene and Bessou (2023) investigated deep learning approaches based on FastText embeddings combined with LSTM networks, reporting improved performance on Algerian Twitter data.

More recently, transformer-based models have shown promising results in Algerian sentiment analysis. (Benmounah et al., 2023b) reported an F1-score of 78.38% and an accuracy of 81.74%. In addition, DziriBERT, introduced by (Abdaoui et al., 2022) and trained on one million Algerian tweets, outperformed larger-scale transformer models on relevant evaluation benchmarks.

Despite these advances, there remains a lack of standardized benchmarking for ADSA models, particularly on platforms such as X. Therefore, this research seeks to address this gap by critically evaluating existing benchmarking efforts and introducing a novel sentiment analysis model tailored to the unique challenges of Arabic and the Algerian dialect. Table 2.3 summarizes key studies in Arabic and dialectal sentiment analysis,

including those focused on the Algerian dialect.

Arabic And Algerian Dialect Datasets

The scarcity of high-quality resources, such as annotated datasets and sentiment lexicons, continues to pose a major challenge for ASA, particularly when addressing dialectal variations (Oueslati et al., 2020). Nonetheless, recent initiatives have led to the development of large-scale Arabic datasets (see Table 2.4). Despite such progress, benchmarking sentiment analysis for the Algerian dialect on social media platforms remains limited due to the scarcity of publicly available annotated datasets (see Table 2.5). To help address this gap, we introduce a new Algerian dialect sentiment corpus, constructed by scraping data from X, which remains a widely used platform for building Arabic dialect corpora (Qi and Shabrina, 2023).

2.8.5 Related Work in Human–Machine Interaction and Sentiment Analysis

Recent research in HMI has increasingly incorporated sentiment analysis and emotion recognition to enable more adaptive, user-centered, and emotionally intelligent systems. By allowing machines to perceive and respond to users' affective states, these approaches aim to improve interaction quality, engagement, and overall user experience across various application domains.

(Anwar et al., 2023) explored the role of sentiment analysis and HMI in recognizing students' emotions within online learning environments. Their study emphasizes the importance of sentiment-aware systems in mitigating the lack of face-to-face interaction in virtual education, demonstrating that emotion-sensitive HMI can enhance student engagement and satisfaction.

(Fu et al., 2025) introduced Multi-HM, a multimodal emotion recognition dataset designed for human–machine consultation scenarios. The dataset comprises 2,000 professionally annotated dialogues spanning multiple HMI domains and integrates textual, acoustic, and visual modalities to model task-oriented emotional behavior. This work highlights the limitations of existing conversational datasets in capturing context-aware emotional dynamics and underscores the need for domain-specific resources to support adaptive and emotionally intelligent HMI systems.

From a broader perspective, (Ghosh, 2023) examined the integration of sentiment analysis within HMI, emphasizing the importance of emotional awareness in enhancing user experience and supporting user-centered system design. The authors discussed key challenges related to emotion detection and response, reviewed multimodal sentiment recognition approaches, and presented examples of emotion-aware interfaces such as embodied conversational agents. The chapter also outlined future research directions and ethical considerations for emotionally intelligent interactive systems.

Table 2.3 Summary of Related Work on Sentiment Analysis for Arabic and Arabic Dialects

Work	Approach	Language	Dataset	Accuracy (%)	Class
(Baniata and Kang, 2024)	MTL MoE	MSA, Dialect	HARD BRAD LABR	84.02 67.89 83.91	5
(Jbel et al., 2024)	ML DL	Moroccan	Social media	92	2
(Alqarni and Rahman, 2023)	CNN BiLSTM	Saudi Arabian	X	92.80	3
(Elhassan et al., 2023a)	DL	MSA, Dialect	HARD LABR	94.54	3
(Jaballi et al., 2022)	ML DL	Tunisian	TUNIZI	88	2
(Farha and Magdy, 2021)	CNN-LSTM	MSA, Dialect	SemEval2017-A ArSAS ASTD	62 92 66	3
(Wazrah and Alhumoud, 2021)	AraBERT SGRU SBi-GRU	MSA	X	90	3
(Abdul-Mageed et al., 2021b)	MARBERT	MSA, Dialect	AraTwitter	F1= 93.4	3
(Ombabi et al., 2020)	CNN-LSTM SVM	MSA	ASTD LABR TwitterASA	90.75	2
(Mohammed and Kora, 2019)	DL	MSA	X	81.31	2
(Abo et al., 2019)	Lexicon-based DT	Sudanese	X	83.5	2
(Alayba et al., 2018)	CNN-LSTM	MSA	Ar-Twitter Ar Health	88.1 94.3	2
(Abdedaiem et al., 2024a)	ML PLM	Algerian	Facebook, YouTube	83.43	3
(Benmounah et al., 2023b)	PLM	Algerian	YouTube	81.7	3
(Mazari and Djeffal, 2022a)	ML DL	Algerian	Facebook X	84.21	2
(Moudjari et al., 2020)	ML, DL	Algerian	YouTube X	76	2
(Mataoui et al., 2016)	Lexicon-based	Algerian	Facebook	79.13	2

Table 2.4 Overview of Sentiment Analysis Datasets for Arabic and Arabic Dialects

Reference	Dataset	Size	Language	Source	Availability
(Alqarni and Rahman, 2023)	-	90,187	Saudi Arabian	X	N/A
(Fourati et al., 2021)	TSAC	17K	Tunisian	SM	Yes
(Alowisheq et al., 2021)	MARSA	61,353	Gulf	X	Yes
(Abdellaoui and Zrigui, 2018)	TEAD	6 million	MD	X	Yes
(Rosenthal et al., 2017)	SemEval 17-Task4	70K	MD	X	Yes
(Alomari et al., 2017)	AGJT	1,800	MD	X	Yes
(Nabil et al., 2015)	ASTD	10,006	MD	X	Yes
(Abdulla et al., 2013)	ArTwitter	1,951	MD	X	Yes
(Aly and Atiya, 2013)	LABR	63,257	MD	Web	Yes

Table 2.5 Overview of Sentiment Analysis Datasets for Algerian Dialect

Reference	Multilingual	Size	Source	Availability
(Abdedaiem et al., 2024a)	MSA, AD	10,087	SM	Yes
(Muhammad et al., 2023)	AD, Arabizi	3,062	AfriSenti	Yes
(Benmounah et al., 2023b)	AD	45K	YouTube	Yes
(Mazari and Djeffal, 2022a)	MSA, French, AD	11,760	SM	N/A
(Rahab et al., 2021)	MSA, French, English	147	newspapers	Yes
(Moudjari et al., 2020)	MSA, AD, Arabizi	9K	X	Yes
(Klouche et al., 2020)	Arabizi, AD, French, English	18,141	Facebook	Yes
(Adouane et al., 2020)	AD, Arabizi	36K	YouTube	N/A
(Abdelli et al., 2019)	MSA, AD	49,864	Facebook	Yes
(Salhi et al., 2019)	MSA, AD	1,510	X	N/A
(Soumeur et al., 2018)	AD, French	25,475	Facebook	N/A
(Mataoui et al., 2016)	AD, French	7,698	Facebook	N/A

Similarly, (Farhadipour et al., 2025) proposed a multimodal framework for emotion recognition and sentiment analysis in conversational contexts by integrating textual, acoustic, facial, and visual information. Leveraging pre-trained models for text and speech alongside specialized architectures for facial and video analysis, the system fuses multimodal features to improve affective state prediction. Experimental results on a realistic multi-party conversational dataset demonstrate the effectiveness of modality fusion for emotion-aware HMI applications.

In contrast to deep and multimodal approaches, (Firdaus et al., 2025) investigated emotion recognition using classical machine learning techniques. Based on speech data from the IEMOCAP dataset, the study applied Naïve Bayes and K-Nearest Neighbor classifiers to identify multiple emotional states. The findings indicate that traditional algorithms can still provide meaningful emotion recognition performance, highlighting their relevance for sentiment-aware and adaptive HMI systems.

Chapter 3

Improving Dialectal Text Classification via Attention-Based Stopword Extraction for Algerian Arabic

3.1 Introduction

Text preprocessing is a crucial step in most NLP pipelines, often involving the removal of stopwords, common words such as conjunctions, prepositions, and auxiliary verbs that carry limited semantic value (Sembok and Abuata, 2025). Effective stopword removal reduces noise, improves computational efficiency, and can enhance the performance of downstream tasks like text classification (Matrane et al., 2023a).

While numerous stopword lists exist for MSA (Sembok and Abuata, 2025), dialectal Arabic remains largely under-resourced. The Algerian dialect poses unique challenges due to its high linguistic variability, rich morphology, and frequent code-switching with French. These characteristics make standard MSA stopword lists insufficient, as they fail to capture the informal expressions and frequent words typical of Algerian dialect, especially in social media texts (Matrane et al., 2025).

Despite growing interest in dialectal Arabic processing (Matrane et al., 2023a), efforts to develop robust, domain-specific stopword lists for Algerian Arabic are limited. Existing lists are often manually created or based solely on frequency heuristics, lacking semantic context and adaptability to modern NLP models (Nassr et al., 2020).

Classical stopword removal methods have limited capacity to identify semantically irrelevant words in context-sensitive tasks (Kuwabara and Suzuki, 2024). In dialectal contexts, frequent words may carry nuanced meanings, such as sarcasm or negation, depending on usage and domain (Kaseb and Farouk, 2023; Alharbi, 2020). Static stopword lists do not adapt to task-specific needs, like sentiment classification, where word importance varies with context.

To address these challenges in low-resource dialectal settings, we propose a model-

driven approach that leverages the attention mechanisms of transformer-based language models. Inspired by prior work that exploits attention for stopword identification (Kuwabara and Suzuki, 2024), we fine-tune MARBERT (Abdul-Mageed et al., 2021b), a transformer pretrained on Arabic and its dialects, for a sentiment classification task and subsequently analyze its attention scores. Tokens that consistently receive low attention across both correct and incorrect predictions are identified as candidates for a context-aware, task-specific stopword list.

The resulting Algerian Arabic stopword list has been made publicly available through the OSF Repository[1] , addressing a critical gap in Arabic NLP resources and supporting the development of more accurate language technologies for underrepresented dialects. To the best of our knowledge, this constitutes the first attention-based stopword list developed specifically for Algerian Arabic.

3.1.1 Classical Stopword List Creation

We applied the TF-IDF metric to the cleaned dataset to identify low-information words. Tokens with TF-IDF scores below a set threshold were selected as candidate stopwords. These were then manually filtered to exclude frequent yet semantically meaningful words. The final list includes 265 refined tokens, including both dialectal and MSA terms common in Algerian Arabic social media.

3.1.2 Attention-based Stopword Extraction

We leverage the attention mechanisms of MARBERT to identify contextually uninformative tokens. After fine-tuning MARBERT for sentiment classification, we extract attention scores for each token from the final transformer layer. Tokens are grouped based on classification outcomes (correct vs. incorrect), and a stopword score is computed as:

$$\text{Score}(w) = \mathbb{E}[\text{Attention}_{incorrect}(w)] - \mathbb{E}[\text{Attention}_{correct}(w)]$$

Higher scores suggest a token may contribute to misclassification, while low, consistent attention indicates low semantic value. We apply a multi-stage filtering process:

- **Attention thresholds:** Tokens within the range [0.0001, 0.03] are retained.
- **Frequency:** Tokens must occur at least 3 times in the dataset.
- **Length:** Tokens fewer than two characters are excluded.
- **Sentiment filtering:** Tokens manually verified against a sentiment lexicon are removed to avoid discarding sentiment-bearing words.

[1] `https://osf.io/ez342/?view_only=bdc2ef75b1ba4fc991b92dd9132eec78`

- **Named Entity Recognition:** Tokens identified as named entities via CAMeL Tools NER (Aldumaykhi et al., 2022) are excluded to preserve semantic meaning.

The resulting list includes 374 tokens, capturing frequent dialectal and MSA terms in Algerian Arabic social media. These stopwords, though common, typically lack sentiment content. Table 3.1 presents a selection of stopwords from our newly generated attention-based stopword list.

Table 3.1 Examples of Algerian Arabic Stopwords with Explanations

Stopword	Explanation
دوك	Means ”now” or ”in a moment”.
زعما	Means ”supposedly” or ”let’s say”.
معليش	Means ”it’s okay” or ”no problem”.
راك	Means ”you are”.
وكي	Means ”when” or ”how”.
تاع	Equivalent to ”of” or a possessive marker.
خويا	Means ”my brother”.

3.2 Experiments and Results

To evaluate the effectiveness of our attention-based stopword list for Algerian dialect sentiment classification, we conducted controlled experiments measuring the impact of different stopword removal strategies on classification performance. We employed both traditional and deep learning models for a comprehensive evaluation.

All experiments utilized the Algerian Dialect Sentiment Corpus (Benmounah et al., 2023a), comprising 45,000 manually annotated social media comments labeled as positive, negative, or neutral. We applied standard preprocessing: lowercasing, punctuation removal, and standard tokenization. A stratified 90/10 train-test split was applied.

3.2.1 Stopword Removal Configurations

We evaluated three distinct configurations:

- Dataset used without removing any stopwords, serving as the baseline.
- Stopwords identified by low TF-IDF scores.
- Our proposed list derived from MARBERT attention analysis.

3.2.2 Classification Models

Two classifiers were employed:

- **SVM**

  Implemented with LinearSVC and a TF-IDF vectorizer. Class imbalance was addressed via class_weight='balanced'.

- **LSTM**

  A recurrent neural network trained on 300-dimensional FastText embeddings for 10 epochs with early stopping. The architecture consisted of an embedding layer (128-dim output), a single LSTM layer (64 units), dropout (rate 0.3), followed by two dense layers (ReLU and softmax activations). The model was optimized with Adam and trained using sparse categorical cross-entropy loss function.

3.2.3 Evaluation Metrics

Given class imbalance, we report macro-averaged Precision, Recall, and F1-score. These compute per-class metrics independently before averaging, ensuring equal weighting regardless of class distribution.

3.2.4 Results

Table 4.6 summarizes the sentiment classification performance of the SVM model under different stopword removal strategies.

The attention-based stopword list yields the highest overall accuracy of 69%, outperforming both the TF-IDF-based list and the baseline (no stopword removal), which both achieve 68% accuracy. This improvement, though modest, indicates that removing stopwords identified via the attention mechanism helps the classifier better focus on informative tokens.

Macro-averaged Precision, Recall, and F1-score metrics also favor the attention-based method, ranging between 0.66–0.67, compared to 0.64–0.66 for TF-IDF and 0.64–0.65 without stopword removal. This reflects a more balanced performance across sentiment classes. The weighted F1-score follows the same trend, with the attention-based approach achieving 0.69, confirming the benefit of stopword removal weighted by class support.

Class-wise analysis shows that the Positive class is consistently well classified across all setups, with precision and recall highest for the attention-based list (Precision: 0.72, Recall: 0.85, F1: 0.78), demonstrating that removing attention-based stopwords does not degrade detection of positive sentiment.

The Negative class gains the most from attention-based removal, improving precision (0.74 vs. 0.73) and recall (0.64 vs. 0.63) over the other methods, resulting in a better F1-score (0.69 vs. 0.68). This suggests that the attention-driven stopword list helps SVM better discriminate negative sentiment tweets.

The Neutral class remains the most challenging, with relatively low precision (0.53–0.55), recall (0.44–0.47), and F1 (0.48–0.51) regardless of stopword removal. This aligns with common difficulties in neutral sentiment detection, where tweets often overlap semantically with positive or negative classes.

Table 3.2 SVM Classification Report with Different Stopword Removal Approaches

Stopword Method	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
Attention-Based	Positive	0.72	0.85	0.78	1891
	Negative	0.74	0.64	0.69	1595
	Neutral	0.55	0.47	0.51	1007
	Accuracy		0.69 (4493 samples)		
	Macro avg	0.67	0.66	0.66	4493
	Weighted avg	0.69	0.69	0.69	4493
TF-IDF	Positive	0.70	0.85	0.77	1891
	Negative	0.73	0.63	0.68	1595
	Neutral	0.53	0.44	0.48	1007
	Accuracy		0.68 (4493 samples)		
	Macro avg	0.66	0.64	0.64	4493
	Weighted avg	0.67	0.68	0.67	4493
No Stopwords	Positive	0.71	0.85	0.77	1891
	Negative	0.73	0.64	0.68	1595
	Neutral	0.53	0.44	0.48	1007
	Accuracy		0.68 (4493 samples)		
	Macro avg	0.66	0.64	0.65	4493
	Weighted avg	0.68	0.68	0.68	4493

Table 3.3 shows the sentiment classification performance of the LSTM model across the three stopword removal configurations.

The LSTM achieves the highest accuracy of 78% using the attention-based stopword list, outperforming both the baseline with no stopword removal (77%) and the TF-IDF-based approach (76%). This confirms that the attention-driven stopword removal enhances the model's ability to capture sentiment-relevant features.

The macro-averaged Precision, Recall, and F1-score are all highest for the attention-based method, consistently around 0.74–0.75, compared to 0.72–0.74 for the other methods. Weighted averages also favor the attention-based list, with F1-score at 0.77, indicating improved balanced performance across classes.

Class-wise, the Positive class is the easiest for the LSTM to detect, with precision and recall highest under the attention-based approach (Precision: 0.79, Recall: 0.84, F1: 0.81). Similarly, the Negative class benefits from attention-based stopword removal, showing improved precision (0.75 vs. 0.72–0.74) and recall (0.71 vs. 0.70), resulting in a better F1-score (0.73).

Unlike in the SVM results, the Neutral class shows substantial improvement with attention-based stopword removal (F1: 0.69), compared to 0.66 and 0.68 in the other two setups. This suggests that the LSTM better leverages the cleaner input after removing attention-identified stopwords to differentiate neutral sentiment.

Table 3.3 LSTM Classification Report with Different Stopword Removal Approaches

Stopword Method	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
Attention-Based	Positive	0.79	0.84	0.81	1891
	Negative	0.75	0.71	0.73	1595
	Neutral	0.70	0.68	0.69	1007
	Accuracy		0.78 (4493 samples)		
	Macro avg	0.75	0.74	0.74	4493
	Weighted avg	0.77	0.78	0.77	4493
TF-IDF	Positive	0.77	0.83	0.80	1891
	Negative	0.72	0.70	0.71	1595
	Neutral	0.67	0.65	0.66	1007
	Accuracy		0.76 (4493 samples)		
	Macro avg	0.72	0.73	0.72	4493
	Weighted avg	0.75	0.76	0.75	4493
No Stopwords	Positive	0.78	0.82	0.80	1891
	Negative	0.74	0.72	0.73	1595
	Neutral	0.69	0.67	0.68	1007
	Accuracy		0.77 (4493 samples)		
	Macro avg	0.74	0.74	0.74	4493
	Weighted avg	0.77	0.77	0.77	4493

3.3 Discussion

The experimental results demonstrate that stopword removal strategies significantly impact sentiment classification performance on the Algerian dialect dataset. Both classical (SVM) and deep learning (LSTM) models benefit from stopword lists tailored through attention-based analysis, outperforming traditional TF-IDF-based removal and the baseline without stopword removal.

As shown in Figure 3.1, the attention-based method achieves the highest classification accuracy for both models: 69% for SVM (vs. 68% with TF-IDF and without removal) and 78% for LSTM (outperforming TF-IDF at 76% and no removal at 77%). While the accuracy improvement for SVM is modest, the gain for LSTM is more notable, confirming the benefit of attention-guided stopword elimination. Macro-averaged F1-scores, which reflect balanced class performance, also favor the attention-based method: 0.66 for SVM and 0.74 for LSTM, compared to ranges of 0.64–0.65 (SVM) and 0.72–0.74 (LSTM) for other

methods. These findings suggest that attention-based stopword removal helps both models better capture sentiment-relevant features by reducing noise from non-informative tokens. This advantage is particularly important in dialectal Arabic, where code-switching, informal usage, and lexical variation are prevalent. By leveraging model-internal attention weights, the stopword list becomes task-specific and context-sensitive, leading to more effective preprocessing than generic frequency-based methods like TF-IDF. Nevertheless, the improvements, while consistent, remain moderate, highlighting the inherent difficulty of sentiment analysis in dialectal Arabic.

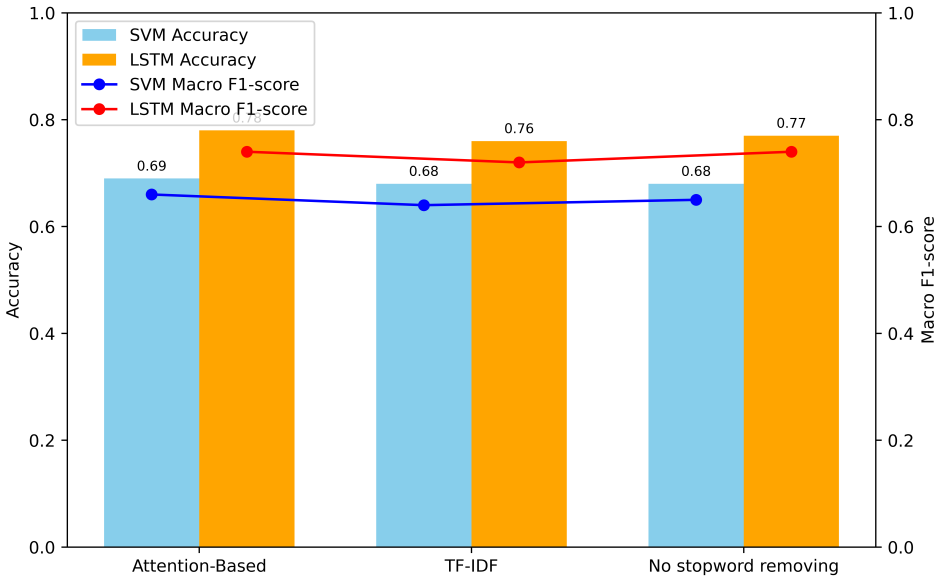


Fig. 3.1 Classification performance comparison (Accuracy and Macro F1-score) of SVM and LSTM models with different stopword removal methods.

3.4 Conclusion and Future Work

We introduced an attention-based stopword removal method leveraging MARBERT to improve sentiment classification in Algerian Arabic. Experiments with both SVM and LSTM demonstrated improved accuracy and more balanced performance compared to TF-IDF-based and no stopword removal approaches. The attention-based method helped models focus on informative words, yielding higher accuracy and more balanced precision and recall across sentiment classes, particularly for challenging categories like negative and neutral. These results confirm the potential of model-based preprocessing to better handle the complexities of dialectal Arabic text.

Future work will aim to further refine attention-based filtering and integrate it with advanced transformer architectures to enhance the robustness of sentiment analysis in low-resource dialectal settings.

Chapter 4

AlgVec: A Word Embedding Model for the Algerian Dialect in Arabic and Arabizi

4.1 Introduction

In recent years, neural word embeddings have emerged as a transformative approach in NLP, capturing semantic and syntactic relationships between words through dense, low-dimensional vectors. Unlike traditional representations such as Bag-of-Words, , or one-hot encoding, which rely on frequency-based statistics and produce sparse, high-dimensional vectors, embeddings enable models to capture meaning more effectively and generalize across semantically related terms (Elhassan et al., 2023b).

Models such as Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), and AraVec (Soliman et al., 2017a) have become essential for a variety of NLP tasks, including machine translation and sentiment analysis. More recently, contextualized embeddings like Bidirectional Encoder Representations from Transformers (BERT) and its Arabic variants, AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021b), have advanced representation capabilities, particularly for morphologically rich and informal languages such as Arabic (Elnagar et al., 2023). These models generate context-sensitive representations that reflect a word's meaning based on its surrounding linguistic context, offering more nuanced semantic understanding than static embeddings. Built on deep learning and self-supervised training over large-scale corpora, they have significantly improved performance across a wide range of NLP tasks (Elnagar et al., 2023).

Despite these advances, most embedding models focus on high-resource languages such as English or MSA, while under-resourced dialects like Algerian Arabic remain largely unsupported (Matrane et al., 2023b). Algerian Arabic is linguistically rich, shaped by historical, social, and regional influences. Strongly impacted by French due to Algeria's colonial past, it features widespread code-switching and lexical borrowing (Amazouz, 2019). It also exhibits considerable regional variation and is often written in Latin script, a practice known as Arabizi, which substitutes Arabic letters with Latin characters and

numerals (e.g., “3” for "ع" or “7” for "ح") (Baert et al., 2020). For example, the phrase *“I don’t know”* may appear in Arabic script as "ما نعرفش" or in Arabizi as “ma na3refch,” “mana3refsh,” or “manarefch.” Regional variants such as "ماعمباليش" or "مانيش عارف" further illustrate the complexity of dialectal expression.

Recent studies have begun addressing the dialectal gap. Resources such as DiaLex (Abdul-Mageed et al., 2021a) and Swan (Bhatia et al., 2025) provide benchmarks and dialect-aware embeddings for various Arabic dialects. However, the Algerian dialect remains underexplored due to the absence of large-scale public datasets and dedicated embedding models (Alosaimi et al., 2024b).

Social media platforms such as X provide a valuable source of dialectal data. In Algeria alone, there were approximately 1.13 million users as of early 2025[1]. These platforms generate millions of informal posts written in both Arabic script and Arabizi. However, such data is challenging to process with standard tools due to its linguistic diversity and lack of normalization. While prior research has demonstrated the potential of X data for Arabic SA (Bouassida and Mezali, 2025), the Algerian dialect remains understudied (Abdedaiem et al., 2024b).

Sentiment Analysis, the task of identifying and classifying emotions or opinions in text, depends heavily on the quality of word representations (Matrane et al., 2023b). In dialectal Arabic, where vocabulary and expressions differ significantly from MSA, generic embeddings often fail to capture relevant nuances. This highlights the importance of training embeddings specifically for dialectal corpora, including support for Arabizi and dialect-specific morphology. To address this gap, we:

- Introduce AlgVec, a suite of word embeddings for the Algerian dialect, trained on over 32 million tokens from X in both Arabic script and Arabizi.
- Evaluate AlgVec through intrinsic tasks, including word similarity, nearest-neighbor retrieval, and a DiaLex-style benchmark, as well as a downstream SA task.
- Propose a combined embedding approach that integrates AlgVec with the transformer-based MARBERT model to enhance performance on dialect-specific benchmarks.
- Release AlgVec and associated resources publicly to support reproducibility and future research in Arabic NLP.

4.2 Research Objectives

The primary objective of this study is to develop word embedding models tailored to the Algerian dialect in both Arabic script and Arabizi. Specifically, the study aims to:

[1] `https://datareportal.com/reports/digital-2025-algeria`

- Train word embedding models that capture semantic and syntactic relationships in both scripts.
- Evaluate the embeddings through intrinsic tasks, including word similarity, nearest-neighbor retrieval, and a DiaLex-style benchmark.
- Assess the effectiveness of the embeddings in a downstream task.

4.3 Building the AlgVec Embeddings

4.3.1 Data Collection

The primary aim of this work is to build robust word embedding models specifically designed for NLP tasks in the Algerian dialect. To achieve this, we constructed a representative corpus using data collected from X, selected for its rich metadata features, particularly geolocation, which allowed us to efficiently identify tweets originating from Algeria. We utilized the X Academic API v2[2], which provides access to up to 10 million tweets per month. Our collection strategy targeted Algerian geolocated tweets (place_country:DZ), excluded retweets to avoid duplication, and spanned the period from 2018 to 2023. In total, we gathered approximately 3 million tweets in the Algerian dialect, covering diverse topics such as politics, health, sports, and daily conversations.

To further strengthen coverage of Arabizi, we incorporated an external Algerian Arabizi corpus (Touileb and Barnes, 2021). This dataset, sourced from publicly available Algerian dialectal resources, was merged with our collected tweets, ensuring representation of both Arabic script and Arabizi. The resulting corpus offers a diverse and representative sample of informal Algerian language use in both scripts. Table 4.1 summarizes the key statistics.

Table 4.1 Statistical summary of the collected Algerian dialect X corpus.

Statistic	Value
Total tweets	3,013,028
Arabic script	62%
Arabizi script	35%
Average tokens per tweet	10.64
Vocabulary size	32,040,423

4.3.2 Data Preprocessing

Before training the embedding models, we applied a series of preprocessing steps to clean and normalize the collected tweets. Given the informal nature of social media content

[2]https://developer.x.com/en/support/x-api/v2

and the linguistic complexity of the Algerian dialect, careful preprocessing was essential to ensure high-quality input.

1. **Language filtering.** Although tweets were geolocated within Algeria, not all were written in the Algerian dialect. We therefore retained only tweets in Arabic script and Arabizi. The `lang` field[3] provided by the X API was used to automatically detect language and to filter out non-textual content (e.g., tweets consisting solely of mentions, hashtags, or media). Tweets marked as `und` (undefined) were discarded to improve data quality.

2. **Data cleaning.** We removed X-specific elements such as user mentions (usernames), hashtags (#tags), emojis, URLs, special characters, punctuation, and extra whitespace.

3. **Normalization.** To standardize variations common in Arabic social media text, we applied:

   a) *Repeated letter reduction:* sequences of more than two consecutive identical letters were reduced (e.g., "ههههه" → "هه").

   b) *Tatweel removal:* elongation characters used for emphasis (e.g., "شكـــرا " → "شكرا").

   c) *Diacritic removal:* short vowels (ً ٌ ٍ َ ُ ِ ّ ْ) were removed.

   d) *Letter standardization:* letter variants were normalized (e.g., "أ", "آ", "إ" → "ا"; "ة" → "ه"; "ى" → "ي").

   e) *Arabizi processing:* all Arabizi text was converted to uppercase to reduce inconsistencies. Digits were removed, except those representing Arabic sounds (e.g., ”3” → "ع", ”7” → "ح").

These steps grouped semantically equivalent words under consistent forms, reduced sparsity, and improved embedding quality.

4.3.3 Embedding Techniques Used in AlgVec

This section describes the techniques employed to build the AlgVec embeddings from the preprocessed data. We used two widely adopted word embedding models: Word2Vec and FastText. These models were chosen for their ability to generate dense vector representations and to effectively handle the informal, noisy text typical of social media and Arabizi.

[3]https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet

Word2Vec Architecture

Word2Vec (Mikolov et al., 2013) is a foundational embedding technique that employs a shallow, two-layer neural network to represent words as dense vectors. It provides two main architectures:

1. **CBOW.** CBOW predicts a target word from its surrounding context. For example, in the Algerian Arabic sentence "الساك عجبني بزاف" ("I really liked the bag"), given the context words ["بزاف", "الساك"], the model attempts to predict the center word "عجبني". CBOW averages the context word vectors to infer the most likely target word. Formally, it maximizes the conditional probability in Eq. (4.1):

$$P(w_t \mid w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+k}) = \frac{\exp\left(v_{w_t}^\top \cdot \frac{1}{2k} \sum_{\substack{-k \le j \le k \\ j \neq 0}} v_{w_{t+j}} \right)}{\sum_{w \in V} \exp\left(v_w^\top \cdot \frac{1}{2k} \sum_{\substack{-k \le j \le k \\ j \neq 0}} v_{w_{t+j}} \right)} \quad (4.1)$$

   Where:

   - w_t: target (center) word, e.g., عجبني
   - w_{t+j}: context words, e.g., الساك, بزاف
   - v_w: embedding vector of word w
   - V: vocabulary
   - k: context window size (e.g., $k = 1$ for two context words)

2. **Skip-gram.** The Skip-gram model predicts surrounding context words given a center word. Using the same example, if the input word is "عجبني", the model predicts its neighbors "الساك" and "بزاف". Formally, it maximizes the conditional probabilities in Eq. (4.2):

$$\prod_{\substack{-k \le j \le k \\ j \neq 0}} P(w_{t+j} \mid w_t) = \prod_{\substack{-k \le j \le k \\ j \neq 0}} \frac{\exp(v_{w_{t+j}}^\top v_{w_t})}{\sum_{w \in V} \exp(v_w^\top v_{w_t})} \quad (4.2)$$

   Where:

   - w_t: input (center) word, e.g., عجبني
   - w_{t+j}: predicted context words, e.g., الساك, بزاف
   - v_w: embedding vector of word w
   - V: vocabulary
   - k: context window size (e.g., $k = 1$)

FastText Architecture

FastText (Bojanowski et al., 2017) extends the Skip-gram architecture by representing each word as a collection of character n-grams. Instead of assigning a single vector to each word, FastText learns embeddings for subword units, which are then summed to obtain the word vector. This approach enables the model to generate embeddings for Out-Of-Vocabulary (OOV) words by composing them from known character n-grams. For example, in the Algerian Arabic word "عجبني" from the sentence "الساك عجبني بزاف", FastText decomposes the word into character n-grams such as "<عج", "عجب", "جبن", "بني", and "ني>", where angle brackets denote the word boundaries.

Given a word w, let G_w denote the set of its character n-grams. The word embedding $\mathbf{z}_w$ is obtained by summing the embeddings of its subword units, as defined in Eq. (4.3):

$$\mathbf{z}_w = \sum_{g \in G_w} \mathbf{z}_g \tag{4.3}$$

FastText adopts a Skip-gram objective with subword embeddings. Given a target word w_t and its context words w_{t+j} within a window of size c, the model maximizes the conditional probability of predicting the context words from the n-grams of the target word. This is expressed in Eq. (4.4):

$$\prod_{\substack{-c \le j \le c \\ j \neq 0}} P(w_{t+j} \mid G_{w_t}) = \prod_{\substack{-c \le j \le c \\ j \neq 0}} \frac{\exp(\mathbf{v}_{w_{t+j}}^{\top} \mathbf{z}_{w_t})}{\sum_{w \in V} \exp(\mathbf{v}_w^{\top} \mathbf{z}_{w_t})} \tag{4.4}$$

Where:

- w_t: the target (center) word, e.g., عجبني
- w_{t+j}: the context words, e.g., الساك and بزاف
- $\mathbf{z}_{w_t}$: subword-based embedding of the target word (sum of n-grams)
- $\mathbf{v}_w$: output word vector of word w
- V: the vocabulary
- c: context window size (here, $c = 1$)

By leveraging subword information, FastText generalizes effectively to morphologically rich and informal languages such as the Algerian dialect.

Model Training and Configuration

To construct the AlgVec embeddings for Algerian Arabic and Arabizi, we trained three static models on a preprocessed corpus of approximately 3 million tweets: Word2Vec-CBOW, Word2Vec-Skip-gram, and FastText-Skip-gram. These models were chosen for

their effectiveness in capturing the semantic and morphological characteristics of dialectal expressions (Elhassan et al., 2023b) frequently found in Algerian social media, as illustrated in Figure 4.1.

All models were trained with a vector size of 300 dimensions, a widely adopted standard in Arabic NLP that balances expressive capacity and computational efficiency (Allahim and Cherif, 2024). The context window size was set to 3 to reflect the short, informal nature of tweets. To reduce noise and highlight meaningful patterns, we applied a minimum word frequency threshold (`min_count`) of 10. Negative sampling with 5 negative examples per positive instance was also employed to enhance stability and efficiency.

For the FastText-Skip-gram model, subword-level learning was enabled using character n-grams ranging from 3 to 6 characters. This feature allows the model to generate embeddings for OOV terms and to better capture the orthographic variability and morphological richness (Bojanowski et al., 2017) inherent in Algerian dialect and Arabizi.

Each AlgVec variant, AlgVec-CBOW, AlgVec-Skip-gram, and AlgVec-FastText, was trained independently on the same corpus under identical hyperparameter settings, ensuring a fair and systematic comparison in downstream tasks.

To enrich the representation of dialectal and informal language, we also explored a hybrid approach that combines static AlgVec embeddings with MARBERT (Abdul-Mageed et al., 2021b), a pre-trained transformer model specialized in Arabic. In this configuration, token-level contextual embeddings from MARBERT (768 dimensions) were concatenated with the averaged static embeddings from each AlgVec model (300 dimensions). The resulting 1068-dimensional sentence representation integrates MARBERT's contextual depth with the dialect-specific lexical knowledge of AlgVec, aiming to improve the robustness of sentiment classification for informal and morphologically diverse text.

4.4 Experimental Evaluation and Results

We conducted a comprehensive evaluation of the AlgVec embeddings using two complementary strategies: intrinsic evaluation and extrinsic evaluation.

4.4.1 Experimental Environment

All experiments were carried out on Google Colab Pro using an NVIDIA Tesla T4 GPU. The implementation was developed in Python with TensorFlow for deep learning models, Hugging Face Transformers for MARBERT, scikit-learn for traditional classifiers, and Gensim for AlgVec training.

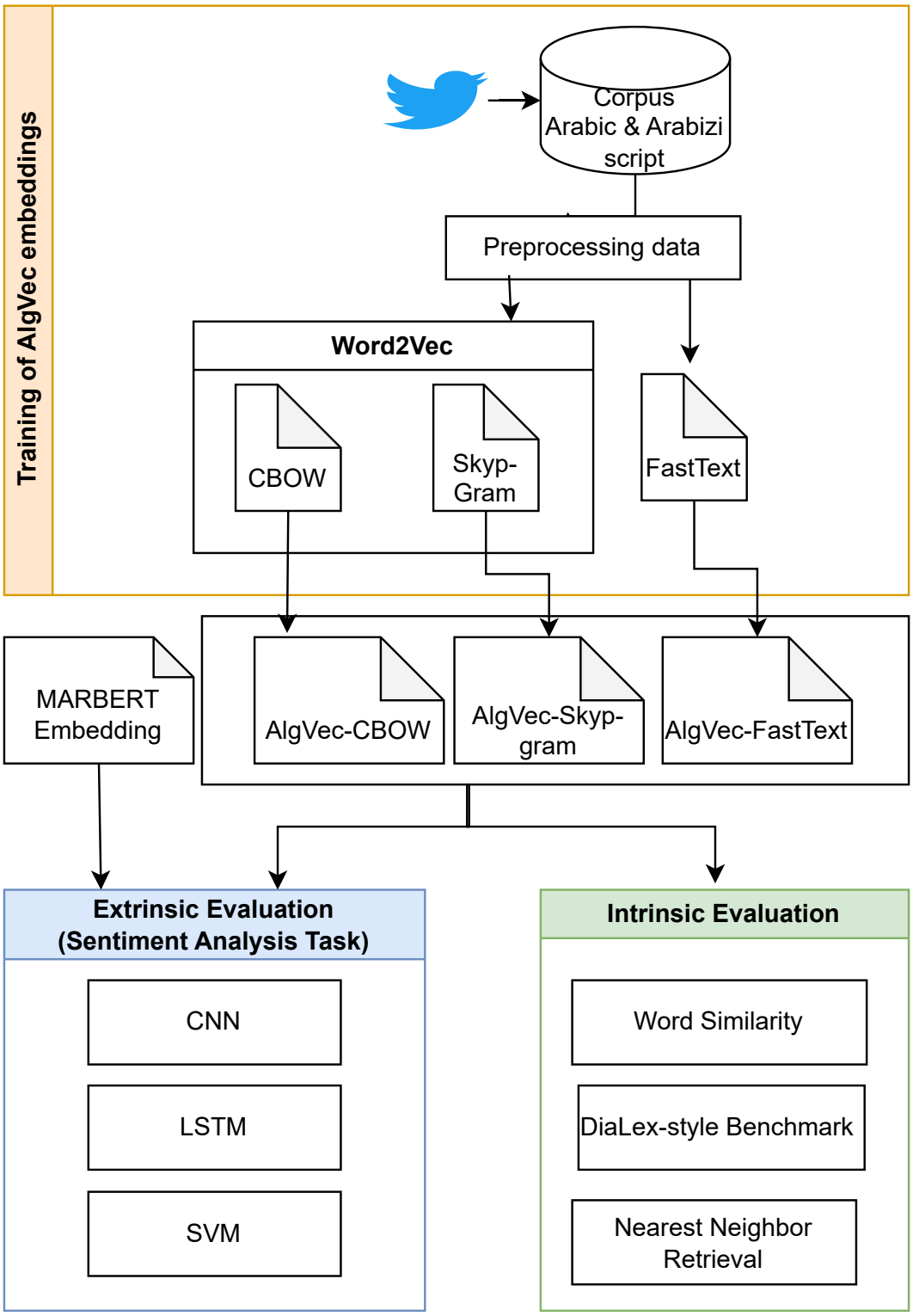


Fig. 4.1 Overview of the AlgVec development pipeline, including data preprocessing, embedding training and evalutaion tasks.

4.4.2 Intrinsic Evaluation

The intrinsic evaluation examined the linguistic properties encoded in AlgVec representations. We generated embeddings using three architectures, FastText, Skip-gram, and CBOW, to compare their strengths and limitations in modeling the Algerian dialect in both Arabic and Arabizi scripts. The evaluation involved three tasks: word similarity, DiaLex-style benchmark relations, and nearest-neighbor retrieval.

Word Similarity

To assess how effectively AlgVec captures semantic similarity, we curated a benchmark of **412** word pairs representative of the Algerian dialect in both Arabic and Arabizi scripts. The pairs were drawn from three sources: (1) high-frequency terms in our collected corpus, (2) dialectal dictionaries[4], and (3) manual curation to ensure coverage of synonyms, antonyms, morphological variants, and unrelated terms. This combination ensured both linguistic diversity and dialectal representativeness.

To ensure reliable and consistent human annotations, a comprehensive guideline document was prepared and distributed to all annotators before the task. The document defined the notion of semantic similarity in the context of Algerian dialect, provided clear rating criteria, and included illustrative examples for each level of similarity on a five-point Likert scale (1 = completely dissimilar, 5 = highly similar). Annotators were instructed to focus on the perceived semantic closeness of words in everyday Algerian usage, especially as used on social media. The guidelines also explained how to handle ambiguous or context-dependent pairs and orthographic variations between Arabic and Arabizi forms.

Three native speakers of Algerian dialect participated in the annotation process. All were linguistically trained and fluent in both writing systems (Arabic and Arabizi), ensuring familiarity with dialectal, phonetic, and orthographic variations. Before beginning the full annotation, the annotators completed a pilot phase with 30 sample word pairs to align interpretations of the similarity scale and resolve any inconsistencies. The pilot results were discussed collectively to refine shared understanding. During the main task, each annotator independently rated all 412 pairs according to the provided scale. Final human similarity scores were obtained by averaging the three ratings.

Inter-rater reliability, measured using Cohen's kappa, was $\kappa = 0.74$, indicating substantial agreement among annotators. For example, the pair (لوطو, طوموبيل), both meaning "car," received an average human score of 4.7, reflecting their near-synonymous use in Algerian Arabic. Such examples demonstrate how the benchmark effectively captures synonymy and dialectal lexical variation.

[4]For Algerian Arabic script, we used the Mo3jam Algerian Arabic Dictionary: `https://en.mo3jam.com/dialect/Algerian/all/`; for Arabizi, we referred to Glosbe: `https://glosbe.com/arq/en`, with subsequent manual validation by native speakers.

We then computed cosine similarities between word-pair vectors across the three AlgVec models and compared them with human judgments using Spearman's rank correlation coefficient (ρ). As shown in Table 4.2, FastText and Skip-gram achieved strong correlations with human ratings ($\rho = 0.85$, $p < 10^{-24}$), while CBOW obtained a lower but still significant correlation ($\rho = 0.75$, $p = 0.037$).

These results indicate that FastText and Skip-gram are particularly effective at modeling semantic relationships in Algerian dialect, including orthographic and phonetic variation across scripts, while CBOW performs better in capturing general contextual similarity.

DiaLex-style Benchmark Evaluation

To provide a structured and linguistically grounded evaluation, we employed the DiaLex benchmark (Abdul-Mageed et al., 2021a), focusing on the Algerian Arabic subset. This benchmark evaluates embeddings based on predefined morphological and semantic categories such as singular–plural, masculine–feminine, antonymy, dual forms, and genitive-to-past transformations.

Specifically, we used the Algerian Arabic resource (DZ.txt[5]), which contains approximately 23,000 word pairs spanning multiple relation types, as shown in Table 4.3. We computed cosine similarity for each pair across the three embedding models, where higher average similarity indicates better alignment with the intended linguistic relation.

As shown in Table 4.3, FastText consistently outperformed the other models across most categories. For instance, in the genitive, past transformation category, FastText scored 0.5022, outperforming Skip-gram (0.4199) and CBOW (0.3440). In the antonymy category, one of the most challenging relations, FastText again achieved the highest score. These results highlight the advantage of subword-based representations in modeling the morphologically rich and orthographically diverse Algerian dialect (Ahmed and Elhadj, 2024).

Nearest Neighbor Retrieval

To further assess the semantic coherence of the embedding space, we performed a nearest neighbor retrieval task. For each selected query word, we retrieved the top-5 most similar words based on cosine similarity. This task illustrates how well the embeddings group semantically or morphologically related words, particularly within Algerian Arabic.

We selected representative query words in both scripts, including "فرحان" (happy), "كرهت" (I hated), and the Arabizi expressions **"nchallah"** (hopefully), **"mlih"** (good, positive), and **"krah"** (hate, negative). As shown in Table 4.4, the models exhibit complementary strengths: CBOW and Skip-gram tend to capture semantic associations, while

[5] https://github.com/UBC-NLP/dialex/blame/main/benchmarks/DZ.txt

Table 4.2 Example word pair similarity scores in Algerian dialect using AlgVec embeddings.

Group	Word 1	Word 2	Human Score	AlgVec-FT	AlgVec-Skip-gram	AlgVec-CBOW
Related	لوطو	طوموبيل	4.7	0.5260	0.5998	0.7017
Related	جميل	رائع	4.8	0.6639	0.6322	0.6645
Opposites	كرهت	فرحان	1.5	0.2249	0.2094	0.1271
Variant Spellings	yahdik	yehdik	5	0.9409	0.8953	0.8839
Unrelated	دار	بحر	1	0.1464	0.1660	0.1231

AlgVec-FT refers to AlgVec-FastText.

Table 4.3 Comparative cosine similarity scores across semantic relations in Algerian dialect embeddings (Skip-gram, CBOW, FastText).

Relation	Nb Word-pairs	Skip-gram	CBOW	FastText	Samples
double	8	**0.5323**	0.4480	0.5060	نتي-نتوما (you–you two)
plural	7,563	0.4870	0.4165	**0.5440**	مسمار-مسامير (nails-nail)
genitive_past_tense	2,983	0.4199	0.3440	**0.5022**	شريان-شرا (bought-buying)
opposite	11,903	0.3490	0.3146	**0.3665**	نجح-دوبل (repeated-succeeded)
man_woman	886	0.5365	0.4843	**0.5629**	عريس-عروس (groom-bride)

FastText excels at retrieving morphological and orthographic variants, including for OOV Arabizi terms.
Results for Arabic script queries:

- For "فرحان", CBOW and Skip-gram retrieved semantically related terms such as "مبسوط" (glad) and "مستمتع" (enjoying), reflecting general semantic understanding. In contrast, FastText returned morphological variants like "فرحانه", "فرحانا", "فرحاني", highlighting the strength of subword modeling.

- For the negative sentiment term "كرهت", CBOW and Skip-gram produced emotionally aligned expressions such as "مليت" (I'm fed up), "تعبت" (I'm tired), and "غاضتني" (It upset me), indicating recognition of sentiment similarity. FastText, however, emphasized close lexical forms such as "كرهتش", "كرهتي", "كرهتني", again demonstrating sensitivity to morphological variation.

Results for Arabizi queries:

- For the seed word **"nchallah"**, FastText retrieved spelling variants such as **"anchallah"**, **"inchallah"**, and **"inchallahe"**, showing its ability to generalize across common Arabizi orthographic differences. In contrast, Skip-gram and CBOW returned semantically related words like **"marbouha"** (blessed) and **"kheir"** (good), which reflect the positive sentiment typically associated with **"nchallah"**.

- For **"mlih"** (good), FastText produced related forms such as **"mliha"** (feminine form), **"3lih"**, and **"ya3tik"** (may God give you), illustrating its ability to model morphologically rich and gendered terms. Meanwhile, CBOW retrieved contextually positive expressions such as **"bezaf"** (a lot), **"fika"** (in you), and **"rak"** (you are). These results suggest that CBOW and Skip-gram capture broader sentiment-related usage, while FastText remains grounded in lexical similarity.

Table 4.4 Top-5 nearest neighbors of sentiment words in Algerian dialect embeddings.

Model	Query Word	Top-5 Nearest Neighbors
FastText	فرحان	فرحاني, الفرحان, فرحانا, وفرحان, فرحانه
Skip-gram	فرحان	وفرحان, فرحانه, فرحاان, مستمتع, فرحانين
CBOW	فرحان	وفرحان, مبسوط, مستمتع, فرحانه, فرحانين
FastText	كرهت	كرهتني, اكرهت, وكرهت, كرهتش, كرهتي
Skip-gram	كرهت	مشفت, كرهنا, وملیت, وكرهت, كرهتني
CBOW	كرهت	ملیت, تعبت, اكره, كرهنا, غاضتني
FastText	nchallah	anchallah, inchallah, inchallahe, challah, machallah
Skip-gram	nchallah	brabi, inchalah, inchaa, marbouha, wladna
CBOW	nchallah	bnchalah, inchalah, bladna, kheir, bark
FastText	mlih	3lih, mliha , 3a , m3akom , ya3tik
Skip-gram	mlih	3adna, tkon , denia , saadan , raho
CBOW	mlih	bezaf , fika , rahou , rak , ghi
FastText	krah	krahna, lbarah, kah, awah, ibrahima
Skip-gram	krah	<OOV>, <OOV>, <OOV>, <OOV>, <OOV>
CBOW	krah	<OOV>, <OOV>, <OOV>, <OOV>, <OOV>

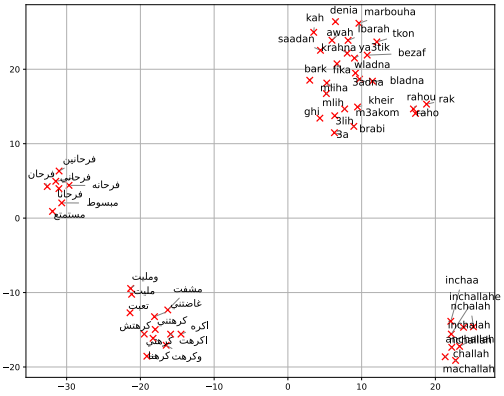
- The analysis of **"krah"** (hate) highlighted the limitations of Skip-gram and CBOW when dealing with OOV items, as neither model returned meaningful neighbors. FastText, however, still produced plausible variations such as **"krahna"** (we hated), **"lbarah"** (yesterday), and **"kah"**, demonstrating robustness to OOV issues. In Table 4.4, these missing outputs are labeled as OOV to clarify the limitation of classical embedding models on rare or irregular Arabizi terms.

We also generated t-SNE plots (Figure 4.2) to visualize how sentiment-related and dialect-specific terms cluster in the embedding spaces.
Overall, the intrinsic evaluations show that FastText excels at modeling morphological and orthographic variation, while Skip-gram and CBOW are more effective at capturing broader semantic associations.

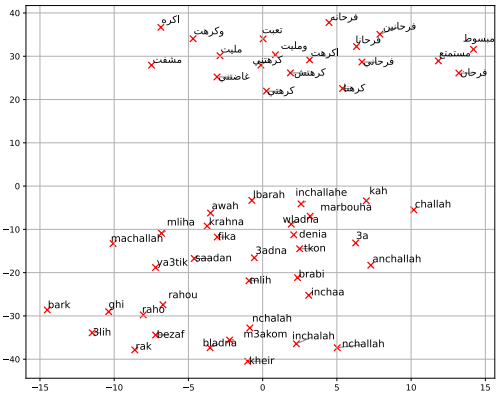
4.4.3 Sentiment Analysis Task

To evaluate the practical utility of AlgVec in real-world NLP applications, we conducted an extrinsic evaluation on a SA task. The goal was to test how effectively the embeddings capture both semantic and affective information, with particular attention to the nuanced and context-dependent expressions of the Algerian dialect.

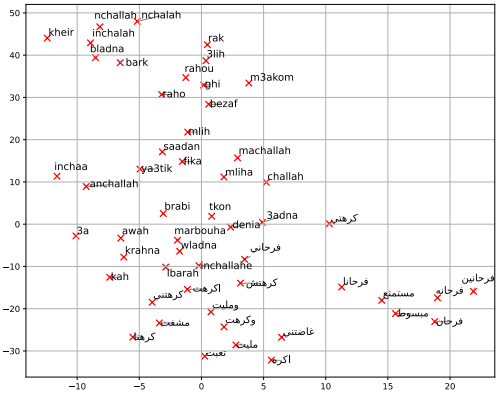
This evaluation demonstrates the generalizability of AlgVec and its ability to handle sentiment-rich, informal dialectal text. The SA pipeline integrates AlgVec with MARBERT, leveraging the strengths of static and contextual representations. To isolate their



(a) FastText



(b) Skip-gram



(c) CBOW

Fig. 4.2 t-SNE visualization of sentiment word embeddings showing clustering of sentiment and dialect-specific terms in Algerian dialect.

respective contributions, we conducted ablation experiments under three configurations: (1) static embeddings only (AlgVec vs. AraVec), (2) contextual embeddings only (MARBERT), and (3) hybrid embeddings combining MARBERT with AlgVec.

Dataset

For the SA experiments, we used the publicly available dataset introduced by Benmounah et al. (2023b), which consists of 45,000 manually annotated Algerian dialect comments collected from social media. Each comment is labeled as positive, negative, or neutral. A key characteristic of this dataset is the class imbalance among sentiment labels. To mitigate the bias caused by this imbalance, we report macro-averaged precision, recall, and F1-score. These metrics calculate performance per class and then average the results, ensuring that minority classes are weighted equally.

To maintain consistency with the original experimental protocol, we followed the same data split defined by the dataset authors: 90% for training and 10% for testing. Prior to model training, the text was preprocessed by removing special characters, punctuation, and emojis, normalizing whitespace, and lowercasing Arabizi text. These steps produced a cleaner and more standardized input, improving the reliability of the models during training and evaluation.

Baseline Experiments: AlgVec Embeddings

To establish a baseline for evaluating our proposed embeddings, we conducted sentiment classification experiments using classical and deep learning models with AlgVec embeddings only. We tested three AlgVec variants, CBOW, Skip-gram, and FastText, and compared them against AraVec (Soliman et al., 2017a), a widely used static embedding model for Arabic. This comparison allowed us to assess the added value of embeddings tailored for the Algerian dialect and Arabizi. For downstream classification, we used three modeling approaches:

- **SVM** A traditional linear classifier using the average of word embeddings for each sentence. This serves as a strong, interpretable baseline.
- **CNN** A 1D convolutional model with 128 filters (kernel size 5), followed by global max pooling, dense layers, dropout, and a softmax output for three-way sentiment classification.
- **LSTM** A recurrent architecture with 128 LSTM units, dropout (0.2), recurrent dropout (0.2), and a softmax classification layer.

For all neural models, AlgVec embeddings were used as fixed weights in the embedding layer, ensuring that performance differences reflect embedding quality rather than additional training. Training was run for 20 epochs with early stopping, a batch size of 64,

and the Adam optimizer with sparse categorical cross-entropy loss. The hyperparameters are summarized in Table 6.5.

Table 4.5 LSTM and CNN model hyperparameters.

Hyperparameter	LSTM Values	CNN Values
Dropout Rate	0.5	0.5
Learning Rate	0.001	0.001
Optimization Algorithm	Adam	Adam
Batch Size	64	64
Number of Epochs	20	20 (EarlyStopping at 10 epochs)

Table 4.6 shows the SVM results. AlgVec-FastText achieved the best performance (74% accuracy, macro F1 = 0.69), followed by AlgVec-CBOW (71%, macro F1 = 0.65) and AlgVec-Skip-gram (70%, macro F1 = 0.62). AraVec produced the lowest scores (64%, macro F1 = 0.55).

Table 4.7 presents the CNN and LSTM results. CNN performance was consistent across embeddings, with accuracies between 72–73%. AlgVec-CBOW and AraVec both reached the highest CNN accuracy (73%, macro F1 = 0.68). In contrast, LSTM models showed stronger performance overall, achieving 74–75% accuracy. AlgVec-CBOW, AlgVec-FastText, and AraVec each reached 75%, with AraVec yielding the best macro precision (0.73) and AlgVec-CBOW the best macro recall (0.70) and F1 (0.71). Weighted F1-scores remained stable (0.73–0.74), reflecting consistent performance across sentiment classes.

Table 4.6 SVM classification results using different word embeddings on the Algerian dialect dataset.

Embedding	Accuracy (%)	Macro Precision	Macro Recall	Macro F1	Weighted F1
AlgVec-CBOW	71	0.68	0.65	0.65	0.70
AlgVec-Skip-gram	70	0.66	0.63	0.62	0.67
AlgVec-FastText	**74**	**0.72**	**0.69**	**0.69**	**0.73**
AraVec	64	0.61	0.57	0.55	0.61

Baseline Experiments: MARBERT Embeddings

To isolate the contribution of contextual embeddings, we conducted experiments using MARBERT without AlgVec. The [CLS] token representation (768 dimensions) was extracted for each sentence and used as input to SVM, CNN, and LSTM classifiers. This setup allows direct comparison with static embeddings (AlgVec, AraVec) and with the hybrid MARBERT+AlgVec models.

Table 4.7 CNN and LSTM classification results using different embeddings on the Algerian dialect dataset.

Model	Embedding	Accuracy (%)	Macro Precision	Macro Recall	Macro F1	Weighted F1
CNN	AlgVec-CBOW	**73**	**0.70**	**0.68**	**0.68**	**0.72**
	AlgVec-Skip-gram	72	0.69	0.67	0.68	0.71
	AlgVec-FastText	72	0.69	0.66	0.66	0.71
	AraVec	**73**	**0.70**	0.67	**0.68**	**0.72**
LSTM	AlgVec-CBOW	**75**	0.72	**0.70**	**0.71**	**0.74**
	AlgVec-Skip-gram	74	0.72	0.69	0.69	0.73
	AlgVec-FastText	**75**	0.72	0.69	0.70	**0.74**
	AraVec	**75**	**0.73**	0.69	0.69	**0.74**

As shown in Table 4.8, MARBERT alone outperformed static embeddings across all classifiers, confirming the benefits of contextual modeling for dialectal SA. However, its performance was slightly below the hybrid models, suggesting that AlgVec provides complementary information by capturing subword morphology and Arabizi-specific lexical variation.

Table 4.8 Classification results using MARBERT-only embeddings on the Algerian dialect dataset.

Model	Embedding	Accuracy (%)	Macro Precision	Macro Recall	Macro F1	Weighted F1
SVM	MARBERT	76.20	0.74	0.73	0.73	0.76
CNN	MARBERT	76.85	0.74	0.74	0.74	0.77
LSTM	MARBERT	**78.42**	**0.76**	**0.75**	**0.75**	**0.78**

Enhanced Experiments: MARBERT + AlgVec Embeddings

For the hybrid setup, we concatenated MARBERT's contextual embeddings (768 dimensions) with the average sentence-level AlgVec embedding (300 dimensions), producing a 1068-dimensional feature vector per sentence. This fusion leverages both dynamic context modeling and fixed semantic representations. The same classification models were used:

- SVM with concatenated MARBERT + AlgVec embeddings.
- CNN and LSTM models adapted for 1068-dimensional input vectors.

Training followed the same configuration as the baselines: 20 epochs, batch size 64, early stopping, and Adam optimizer with sparse categorical cross-entropy. Table 4.9 shows that SVM achieved its best accuracy (78.54%) with MARBERT + AlgVec-FastText, along with the highest macro precision, recall, and F1 (all 0.76). This highlights the benefit of combining MARBERT's contextual understanding with FastText's subword-aware representations, particularly for Algerian dialect and Arabizi. Table 4.10 presents CNN and LSTM results. LSTM consistently outperformed CNN, with MARBERT +

AlgVec-FastText yielding the best scores (79.19% accuracy, macro precision 0.77, macro F1 0.76). CNN models performed worse, but still surpassed static-only embeddings.

Overall, the hybrid approach effectively integrates local contextual dependencies (MARBERT) and morphological richness (AlgVec), making it well-suited for SA in resource-scarce dialectal settings. AlgVec-FastText consistently outperformed CBOW and Skip-gram, underscoring the importance of subword modeling for handling non-standard orthography in Algerian Arabic. Among classifiers, LSTM achieved the strongest performance, while CNNs underperformed when combined with MARBERT.

Table 4.9 SVM classification results using MARBERT + AlgVec embeddings on the Algerian dialect dataset.

Embedding	Accuracy (%)	Macro Precision	Macro Recall	Macro F1	Weighted F1
AlgVec-CBOW	77.59	0.75	0.75	0.75	0.78
AlgVec-Skip-gram	77.57	0.75	0.75	0.75	0.78
AlgVec-FastText	**78.54**	**0.76**	**0.76**	**0.76**	**0.79**

Table 4.10 CNN and LSTM results using MARBERT + AlgVec embeddings on the Algerian dialect dataset.

Model	Embedding	Accuracy (%)	Macro Precision	Macro Recall	Macro F1	Weighted F1
CNN	AlgVec-CBOW	56.71	0.38	0.47	0.42	0.49
	AlgVec-Skip-gram	**60.94**	**0.40**	**0.51**	**0.45**	**0.53**
	AlgVec-FastText	59.14	**0.40**	0.50	0.44	0.52
LSTM	AlgVec-CBOW	77.36	0.75	0.73	0.73	0.77
	AlgVec-Skip-gram	79.00	0.77	0.75	0.76	0.79
	AlgVec-FastText	**79.19**	**0.77**	0.75	**0.76**	**0.79**

4.5 Discussion

This section synthesizes the key findings from both intrinsic and extrinsic evaluations of the proposed AlgVec embeddings, including their integration with SVM and deep learning models (CNN and LSTM) for SA of Algerian dialect texts. The experiments covered standalone AlgVec variants, hybrid combinations with the transformer-based MARBERT, and comparisons with widely used Arabic embeddings such as AraVec to assess the relative advantages of AlgVec for dialect processing.

4.5.1 Intrinsic tasks

Spearman rank correlation analysis of the word similarity task showed that FastText and Skip-gram embeddings achieved strong alignment with human-annotated scores ($\rho = 0.85$, $p1.40\times10^{2}$). CBOW also demonstrated good correlation ($\rho = 0.75$, $p0.037$) but was less

consistent in mirroring human semantic judgments. These results indicate that FastText and Skip-gram are particularly effective in capturing semantic similarity in Algerian dialect, closely reflecting human intuition.

On the DiaLex-style benchmark, FastText consistently achieved the highest cosine similarity scores, especially in genitive–past (0.5022) and antonymy categories. Its subword-based design allows it to capture morphological richness and variation, handle OOV words, and adapt to non-standard spelling. Skip-gram and CBOW showed lower and more variable performance, with CBOW weakest in complex relations such as antonyms. Skip-gram performed moderately well but was limited by its reliance on full-word co-occurrence patterns.

Nearest-neighbor analysis further highlighted key differences. CBOW and Skip-gram retrieved semantically or sentimentally related terms, reflecting their strength in capturing contextual semantics. In contrast, FastText retrieved morphological variants, demonstrating its sensitivity to subword-level patterns. This divergence reflects a complementary relationship: FastText excels in modeling morphological variation, while CBOW and Skip-gram are better at capturing higher-level semantic and affective meaning.

Overall, FastText is well-suited for handling informal, noisy, and orthographically variable Arabizi, whereas CBOW and Skip-gram offer advantages in sentiment-related tasks where broader semantic relationships are crucial.

4.5.2 Extrinsic tasks

The SA experiments highlight the practical utility of AlgVec for dialectal Arabic NLP. As shown in Figure 4.3, AlgVec-FastText consistently outperformed other variants across traditional and hybrid architectures. This robustness stems from FastText's ability to capture lexical variation, morphological richness, and orthographic inconsistencies. In contrast, CBOW and Skip-gram were competitive but less generalizable. AraVec, though widely used for MSA, performed poorly in this context, underscoring the limitations of generic embeddings for dialectal applications.

Hybrid embeddings combining MARBERT with AlgVec further improved performance, with MARBERT + FastText yielding the best results. These gains confirm that integrating contextualized models with dialect-specific static embeddings produces richer and more discriminative sentence representations.

Classifier architecture also played a key role. LSTM consistently outperformed CNN, owing to its ability to capture sequential dependencies and long-range context, both essential for informal, context-rich dialectal text. CNNs, while effective for local features, struggled with high-dimensional contextual embeddings such as MARBERT. Their convolution and pooling operations likely discarded important dependencies, leading to weaker performance. Moreover, CNNs generally require larger training datasets, and the relatively modest dataset size may have limited their effectiveness (Zhang and You, 2021).

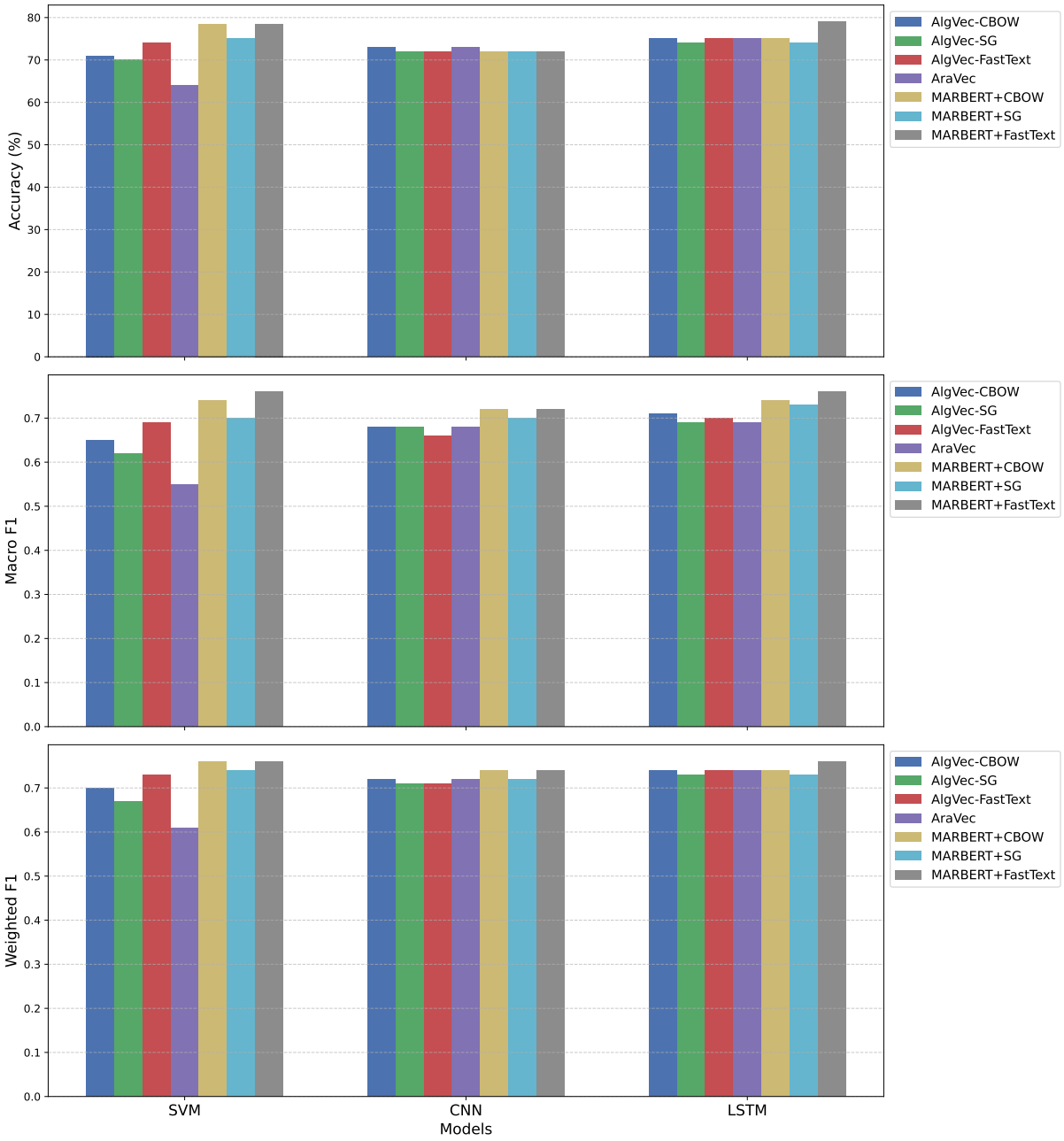


Fig. 4.3 Performance comparison of embedding models (AlgVec-CBOW, AlgVec-Skip-gram, AlgVec-FastText, AraVec, MARBERT with AlgVec variants) on SA using SVM, CNN, and LSTM classifiers.

Future improvements could include attention-augmented CNNs, multi-channel architectures, or joint fine-tuning with MARBERT. Data augmentation may also help CNNs perform more competitively (Zhang and You, 2021).

Overall, the findings show that no single embedding or architecture is universally superior. The most effective results arise from leveraging the complementary strengths of FastText's subword modeling and MARBERT's contextual depth. This synergy is particularly valuable in low-resource, dialectal settings, where both morphological robustness and contextual understanding are crucial. Another important finding is the adaptability of our models: they can be fine-tuned on relatively small amounts of labeled data, enabling rapid domain adaptation without retraining from scratch. This makes them promising for deployment across different platforms beyond X, such as other social media environments.

Qualitative Error Analysis

To better understand model limitations, we analyzed misclassified tweets and identified common patterns. Tweets mixing Arabic script, Arabizi, and French were often misinterpreted, as the models struggled to integrate meaning across scripts. Tweets with implied or subtle sentiment, especially neutral comments with slight emotional cues, were difficult to classify. Rare dialectal words and new Arabizi spellings also caused errors, as static embeddings could not represent them effectively. Table 6.10 presents examples of these misclassifications. These findings point to the need for more context-aware models, sarcasm detection mechanisms, and richer dialectal resources.

4.6 Conclusion

This study introduced AlgVec, a suite of word embeddings specifically designed for the Algerian dialect in both Arabic script and Arabizi. Intrinsic evaluations showed that the different embedding methods capture complementary linguistic features: FastText is particularly effective for handling morphological variation and OOV terms common in informal Arabizi, while CBOW and Skip-gram are better at capturing semantic and sentiment-related context. Extrinsic evaluations on sentiment classification using SVM, CNN, and LSTM models further demonstrated that AlgVec embeddings perform competitively, especially when combined with transformer-based embeddings such as MARBERT. These findings confirm the value of dialect-specific embeddings in enhancing NLP tasks for low-resource Arabic dialects. Overall, AlgVec highlights the necessity of tailored language resources that address the unique traits of dialectal and non-standard Arabic, and provides a foundation for advancing Arabic NLP in applications such as social media analysis, human-computer interaction, and sentiment-aware systems targeting North African dialects.

Future work will expand AlgVec to cover additional North African dialects and improve robustness to code-switching and mixed-script text. We also plan to evaluate the models on other social media platforms, particularly YouTube. This direction aligns with Shevtsov et al. (2023), who found that although topics may be similar across platforms, differences in interaction styles and sentiment expression are substantial. The longer, more varied linguistic style of YouTube in particular poses challenges for models trained exclusively on X data. To address cross-platform adaptation, our hybrid MARBERT + AlgVec-FastText models can be fine-tuned with relatively small amounts of labeled data from new platforms, enabling rapid updating without retraining from scratch. By addressing these challenges, we aim to enhance the generalizability and adaptability of SA models across dialects and platforms.

Table 4.11 Examples of Misclassified Tweets

Tweet	True Label	Predicted Label	Note
هههههه خلاص والله ماقدرت بالظحك رووووعة رووووعة "Hahaha, enough, I couldn't stop laughing, amazing, amazing"	Positive	Negative	Despite the clear positive sentiment, the classifier predicted negative. Repetition of "رووووعة" (amazing) indicated positivity, but excessive laughter tokens "هههه" and informal phrasing introduced noise.
مصنع 04 مليار لازم يا كلو معاك 004 مليار "A 40-billion factory, but they need to take 400 billion with it"	Negative	Neutral	Sarcasm criticizing corruption; sarcasm is hard for models to detect, leading to a neutral label.
lairetam dn3 hcw bas7 al3 am egatnom md5t tnmlmrn orb ko s ti tuB fzzzb lb ri5 t9w le a3m dizt w ml3tt "It depends on the material you have, normally you can do the editing better with practice. But it's ok bro, you learn and improve with time"	Neutral	Negative	The statement is mostly neutral, but the model misinterpreted "ma 5ir bl bzzzf" as criticism, leaning negative.
yawe kiwe "nonsense" or "what is this?"	Negative	Neutral	Short idiomatic slang implying annoyance or rejection. Lack of context and rarity led to neutral classification.
مهابل قراف bonne continuation "Crazy/funny, keep it up"	Positive	Negative	Although "مهابل قراف" literally means "crazy", in this context it conveys humor and encouragement, reinforced by "bonne continuation". The classifier misread the literal meaning and failed to handle code-switching.

Chapter 5

Self-Supervised Transfer Learning for Arabizi Sentiment Analysis

5.1 Introduction

Arabizi is a widely used informal writing system in which Arabic dialects are transcribed using Latin characters and numerals. It is particularly prevalent on social media platforms in North African countries, including Algeria (Fourati et al., 2021). Despite its extensive use in user-generated content, Arabizi remains a challenging and underexplored variant for NLP systems. Its spelling is highly inconsistent, it lacks standardized grammatical conventions, and it is severely underrepresented in annotated datasets (Baert et al., 2020). These characteristics substantially hinder the development of robust NLP models, especially for downstream tasks such as sentiment analysis.

In the Algerian context, Arabizi exhibits extreme orthographic variability. For example, the expression conveying strong positive sentiment, equivalent to "I like it a lot," may appear as *3jbni bzaaf*, *3jebni bzaf*, or *3ajbni bzaaf*, while the word *sa7abti* (my friend) may also be written as *sahebti*. Although these variants convey identical meanings, their divergent surface forms complicate tokenization, embedding generation, and model generalization (Mandalam et al., 2024). As a result, models trained on limited Arabizi data often fail to generalize effectively to unseen spelling variations.

Most existing pretrained language models for Arabic sentiment analysis, including MARBERT (Abdul-Mageed et al., 2021b), are optimized for Arabic script and therefore exhibit limited effectiveness when applied to Arabizi text (Mandalam et al., 2024). Such models struggle to capture transliterated dialectal vocabulary and the digit–letter substitutions characteristic of Arabizi. While Arabizi-specific models have been proposed, such as BAERT (Baert et al., 2020), they are primarily trained on Egyptian Arabizi and do not adequately reflect the linguistic characteristics of North African variants. In contrast, Algerian Arabizi shares stronger lexical and orthographic similarities with Moroccan Arabizi, making models pretrained on Moroccan data more suitable candidates for transfer

learning.

Motivated by these observations, this chapter investigates sentiment analysis for Algerian Arabizi through transfer learning from a closely related dialect. We build upon DarijaBERT-Arabizi (Gaanoun et al., 2024), a pretrained language model for Moroccan Arabizi, and adapt it to Algerian Arabizi using self-supervised learning objectives. Specifically, we refine the encoder through masked language modeling and contrastive learning to better capture orthographic variation, digit–letter substitutions, and informal contextual usage. The adapted representations are then integrated into a hybrid deep learning architecture and fine-tuned on a labeled dataset of Algerian Arabizi tweets for sentiment classification.

The main contributions of this chapter are threefold:

- We propose a transfer learning framework that adapts a pretrained Arabizi language model to Algerian Arabizi using self-supervised objectives.
- We design a hybrid deep learning architecture that combines contextualized Arabizi representations with sequential modeling and self-attention for sentiment classification.
- We provide a comprehensive evaluation demonstrating that the proposed approach outperforms classical machine learning baselines and neural models relying on static embeddings.

5.2 Proposed Architecture

This section describes the overall architecture and transfer learning strategy adopted in this study. We build upon DarijaBERT-Arabizi, a pretrained language model for Moroccan Arabizi, selected because its linguistic patterns and orthographic conventions are closely aligned with those of Algerian Arabizi. The encoder is further refined through self-supervised learning objectives, namely masked language modeling (MLM) and contrastive learning, to enhance its capacity to model the informal spelling, character substitutions, and digit–letter combinations that characterize Arabizi text.

After this self-supervised adaptation, the refined encoder is integrated into a hybrid deep learning architecture comprising:

1. Contextualized embeddings from the adapted DarijaBERT-Arabizi encoder,
2. A bidirectional LSTM for capturing sequential dependencies, and
3. A self-attention mechanism to highlight sentiment-bearing tokens.

Finally, a dense classification layer predicts the sentiment polarity (positive, negative, or neutral). The overall framework is illustrated in Figure 6.2.

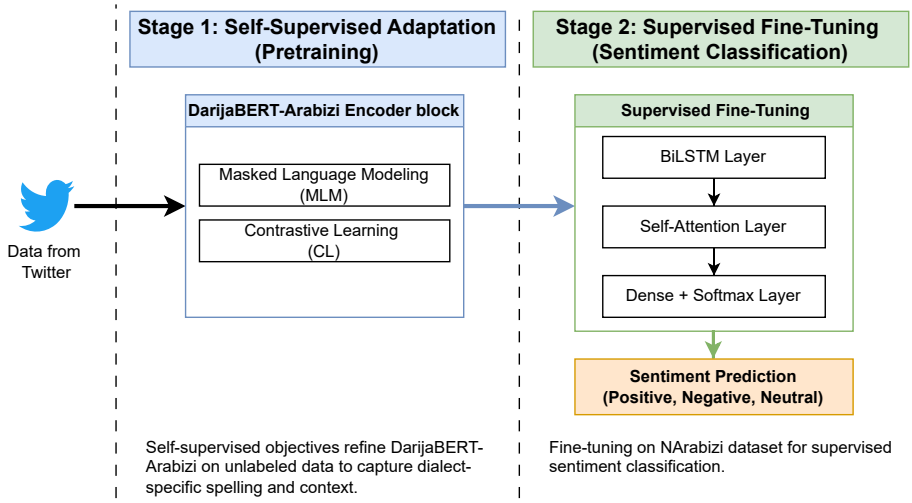


Fig. 5.1 Overview of the proposed hybrid framework for Arabizi sentiment analysis

5.3 Experiments

5.3.1 Dataset Preparation

We use the NArabizi dataset Touileb (2022), a publicly available collection of Algerian Arabizi tweets annotated for sentiment analysis. The corpus contains 1,288 sentences across four sentiment classes: positive (29.87%), negative (27.52%), neutral (18.14%), and mixed (24.47%). Since the *mixed* class reflects ambiguous sentiment that does not align with our classification objective, it was excluded, resulting in a final dataset of 972 sentences.

Preprocessing involved light normalization, including removal of extra whitespace, correction of detached punctuation, and consistent tokenization. Crucially, we preserved spelling variability and digit–letter substitutions (e.g., *bazaf*, *bezzef*), which are characteristic of Arabizi and essential for realistic model training.

5.3.2 Fine-Tuning for Sentiment Analysis

The refined encoder is fine-tuned on the NArabizi dataset for sentiment classification. Fine-tuning is performed end-to-end using categorical cross-entropy as the loss function and the Adam optimizer with a learning rate of 2×10^{-5}. Training is conducted for 10 epochs with a batch size of 16, and early stopping is applied based on validation loss to prevent overfitting. Dropout regularization (rate = 0.3) is applied during training to enhance generalization. This configuration balances model complexity with the relatively small size of NArabizi, ensuring stable and effective learning.

5.3.3 Evaluation Metrics

We evaluate models using **accuracy** and **macro F1-score**. Accuracy offers a general measure of overall performance and allows comparison with prior work. Since the NArabizi dataset is imbalanced, macro F1 is more appropriate as it assigns equal weight to each class, ensuring fair evaluation of minority classes (neutral in our case).

5.3.4 Baseline Comparisons

To evaluate the effectiveness of our proposed framework, we compare it against three categories of baselines:

- **Traditional classifier:** SVM trained with TF-IDF features, representing a strong non-neural baseline.
- **Logistic Regression:** Another classical classifier using TF-IDF features for a fair comparison with SVM.
- **Ensemble:** A soft-voting ensemble of SVM and Logistic Regression, designed to combine the strengths of both models.
- **Pretrained embeddings:** Although ArabiziVec embeddings were specifically developed for Arabizi, they are not publicly available. As an alternative, we use FastText embeddings Bojanowski et al. (2017) combined with a BiLSTM, since their subword modeling capability allows them to capture spelling variations and digit–letter substitutions characteristic of Arabizi.

5.4 Results and Discussion

The experimental results demonstrate that the proposed transfer learning framework is highly effective for sentiment analysis in Algerian Arabizi, significantly outperforming all baseline approaches. As shown in Table 5.1, the proposed model achieves an accuracy of 0.816 and a macro F1-score of 0.808, highlighting the benefit of combining a pretrained Arabizi encoder with self-supervised adaptation and contextual modeling.

Classical machine learning baselines based on TF-IDF representations, including SVM, Logistic Regression, and their ensemble, achieve only moderate performance, with macro F1-scores ranging from 0.532 to 0.574. These results reflect the inherent limitations of shallow lexical representations when applied to Arabizi text, where extensive spelling variation and informal orthography prevent effective feature generalization. The lack of performance improvement from the ensemble model further suggests that TF-IDF-based classifiers capture largely overlapping information and fail to model contextual dependencies.

Neural models based on static word embeddings also show limited effectiveness. The BiLSTM trained with FastText embeddings achieves a macro F1-score of only 0.483, despite FastText's ability to model subword information. While subword modeling partially alleviates spelling variation, it remains insufficient for capturing higher-level semantic and contextual cues required for sentiment disambiguation in noisy dialectal data. This result emphasizes the necessity of contextualized representations for Arabizi sentiment analysis.

In contrast, the proposed framework benefits from transfer learning across closely related dialects and from self-supervised refinement on Arabizi-specific objectives. The ablation study in Table 5.2 confirms the incremental contribution of each component. Masked language modeling improves performance by enabling the encoder to better model orthographic and lexical variation, while contrastive learning further enhances robustness by encouraging discrimination between semantically distinct instances. The integration of BiLSTM captures sequential dependencies, and the self-attention mechanism allows the model to focus on sentiment-bearing tokens, resulting in the highest overall performance.

Although the NArabizi dataset is relatively small, the consistent improvements observed across all evaluation metrics and the systematic gains demonstrated by the ablation study indicate that the proposed approach generalizes effectively within this low-resource setting. Rather than relying on large annotated corpora, the framework leverages self-supervised learning and cross-dialect transfer to mitigate data scarcity, which is a critical requirement for underrepresented dialects such as Algerian Arabizi.

Overall, these findings confirm that transfer learning from linguistically related dialects, combined with self-supervised adaptation, provides a viable and effective strategy for sentiment analysis in Arabizi and other low-resource dialectal variants.

5.5 Conclusion

This work addressed the challenge of sentiment analysis for Algerian Arabizi, a dialectal and highly informal writing style widely used on social media. We proposed a hybrid deep learning framework that leverages DarijaBERT-Arabizi as a contextual encoder, refined through self-supervised learning, and integrates BiLSTM with a self-attention mechanism for sentiment classification.

Experiments on the NArabizi dataset demonstrated that our framework significantly outperforms classical machine learning baselines using TF-IDF features, as well as neural models based on static FastText embeddings. These findings confirm that transfer learning across closely related dialects can mitigate the challenge of limited annotated resources and enable robust sentiment analysis in low-resource settings.

The future work will focus on pretraining the model on larger unlabeled Algerian Arabizi corpora, exploring multilingual transfer strategies, and extending the proposed framework to additional NLP tasks in dialectal Arabic.

Table 5.1 Summary of Experimental Results on NArabizi Dataset

Model	Representation	Accuracy	F1-score (Macro)
Proposed framework	DarijaBERT-Arabizi + BiLSTM + Attention	0.816	0.808
SVM	TF-IDF	0.622	0.532
Logistic Regression	TF-IDF	0.643	0.574
Ensemble (SVM + LR)	TF-IDF	0.643	0.574
BiLSTM	FastText embeddings	0.483	0.483

Table 5.2 Ablation study illustrating the contribution of each component to overall performance (measured by Macro F1).

Model Variant	Macro F1
DarijaBERT-Arabizi (baseline)	0.752
+ MLM	0.774
+ MLM + Contrastive	0.788
+ MLM + Contrastive + BiLSTM	0.799
Full Model (+ Attention)	**0.808**

Chapter 6

Benchmarking Sentiment Analysis of Algerian Arabic Dialect on X (Twitter)

6.1 Introduction

X has emerged as a central platform for expressing opinions, emotions, and public attitudes toward social, political, and economic issues (Hajirahimova and Ismayilova, 2022). The platform produces large volumes of short, informal, and rapidly evolving textual content, making it a valuable yet challenging resource for sentiment analysis (SA) research (Qi and Shabrina, 2023). Its real-time nature and strong association with public discourse have positioned X as a primary benchmark for evaluating sentiment analysis systems across languages and domains.

Sentiment analysis is a core task in NLP that aims to automatically identify the polarity of a text as positive, negative, or neutral (Kumar et al., 2025). Early work relied on classical machine learning approaches, such as SVM and NB, combined with handcrafted features and statistical representations. More recent advances in deep learning and transformer-based models, particularly BERT and its variants, have substantially improved sentiment classification performance by leveraging contextualized representations and self-attention mechanisms (Shi and Agrawal, 2025; Bouassida and Mezali, 2025). These models have proven especially effective for informal and noisy social media data through transfer learning.

Arabic sentiment analysis has attracted increasing attention due to the rapid growth of Arabic content on social media platforms. As of 2025, Arabic ranks among the five most frequently used languages on X [1]. However, the majority of existing Arabic NLP resources focus on MSA, despite the fact that social media communication is dominated by dialectal varieties (Oueslati et al., 2020). Arabic dialects exhibit substantial lexical, morphological, and syntactic variation, which limits the effectiveness of MSA-centric models and highlights the need for dialect-aware sentiment analysis frameworks.

[1] https://semiocast.com/top-languages-on-twitter-stats/

While several studies have addressed sentiment analysis in Arabic dialects such as Saudi (Alosaimi et al., 2024a), Egyptian (Sherif and Sabty, 2024), Moroccan (Matrane et al., 2024; Ouahabi et al., 2023), and Tunisian (Jaballi et al., 2022) varieties, the Algerian dialect remains comparatively underexplored. Algerian Arabic presents additional challenges due to extensive code-switching with French, frequent use of Arabizi, phonetic variability, and the absence of standardized orthography (Rahmoun-Mrabet, 2018; Benrabah, 2013). Existing studies on Algerian sentiment analysis are limited in scale and lack standardized benchmarks, making systematic comparison and reproducibility difficult.

Beyond dialectal variation, cross-domain generalization remains a major challenge in sentiment analysis. Models trained on a specific platform often fail to generalize due to differences in discourse style, text length, and sentiment expression patterns across platforms such as X and YouTube (Shevtsov et al., 2023; Alahmadi et al., 2025). Establishing robust benchmarks that evaluate both in-domain performance and cross-domain transfer is therefore essential.

This chapter addresses these gaps by presenting a comprehensive benchmark for sentiment analysis in the Algerian dialect on X. We introduce a large-scale annotated dataset, design a dialect-aware preprocessing pipeline, and systematically evaluate traditional machine learning models, deep learning architectures, and transformer-based approaches. In addition, we assess the cross-domain robustness of the best-performing model on external datasets. The contributions of this chapter are summarized as follows:

6.2 Methodology

6.2.1 Data Collection and Annotation

To construct our dataset, we utilized the X Academic API v2[2], which allows access to up to 10 million tweets per month. We applied filtering criteria to ensure relevance to Algerian discourse. Specifically, tweets were collected using geolocation filters (place_country:DZ) to target content originating from Algeria; retweets were excluded to reduce redundancy. This process resulted in a dataset of 2.5 million tweets spanning the years 2018 to 2023. The tweets covered a broad range of topics, including politics, health, and sports (see Figure 6.1 for topic distribution).

Due to the linguistic complexities of the Algerian dialect, including code-switching and non-standard spellings, we ensured high-quality data annotation to support effective sentiment analysis (Almuqren et al., 2017). To this end, we manually annotated a dataset of 18,589 tweets, focusing exclusively on content written in Algerian dialect or Modern Standard Arabic (MSA). The resulting dataset was named **"X for Algerian Dialect**

[2]https://developer.x.com/en/docs/tutorials/step-by-step-guide-to-making-your-first-request-to-the-twitter-api-v2

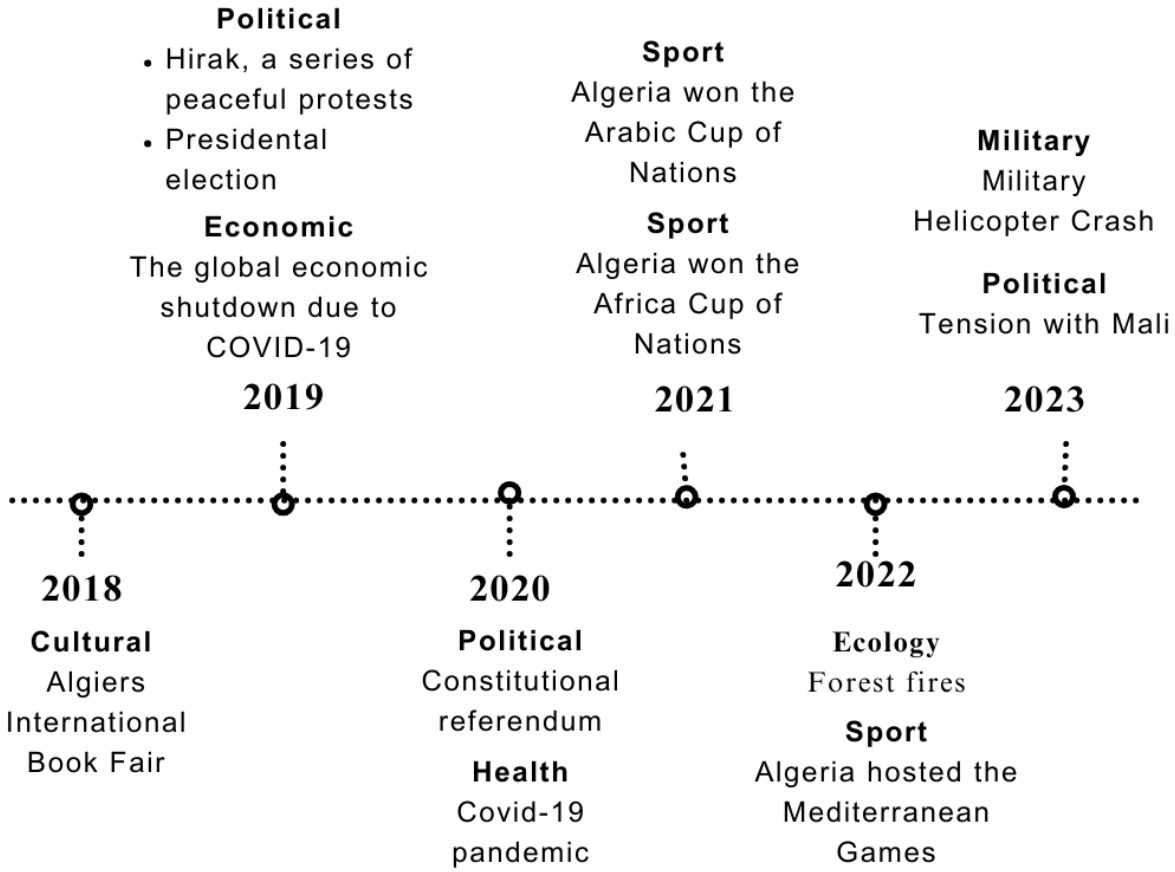


Fig. 6.1 Important events that have occurred in Algeria between 2018 and 2023

Sentiment Analysis" (X4ADSA). Table 6.1 presents illustrative examples, associated sentiment labels, and key linguistic features.

Annotation was carried out by seven native Algerian speakers using a three-class sentiment scheme: positive, negative, and neutral. To maintain consistency throughout the process, we developed comprehensive guidelines that addressed complex cases such as sarcasm, mixed or ambiguous sentiment, code-switching, and Arabizi usage (see Appendix A for details and examples). Each tweet was independently labeled by three annotators using shared Microsoft Excel sheets.

6.2.2 Data Preprocessing

To ensure data integrity and improve the performance of our sentiment analysis system, we designed a multi-phase preprocessing pipeline aligned with the best practices in Arabic social media analysis (Almutairi et al., 2024). At each stage, empty and duplicate entries were systematically detected and removed to maintain the overall quality of the dataset.

Table 6.1 Annotated Examples from the Algerian Dialect Dataset Illustrating Linguistic Challenges and Sentiment Ambiguities

Case Type	Language	Tweet	Sentiment Class
Sarcasm	MSA + AD	أنا نقلك روحي ترقدي واحلمي بالتحليق في الجزائر	Negative
Arabizi	AD + FR	4h du sbah ça l'air fuité	Neutral
Literal Statement	MSA	الجزائر ستفوز مرة آخرا 🏆	Positive
Irony	AD	راهم يوجدوا للريفيون.. هههههه	Negative
Code-switching + Arabizi	MSA + EN	Go to the devil! لا نريدك في أرض الشهداء	Negative
Code-switching + Arabizi	AD + FR	انقطاعات الجزائر شبعتونا خلايع تقول راكم تصدقو conxion علينا كونا نعسو الماء اذا جا ولينا نعس dyalkom tanit grave	Negative

Phase 1: Data Filtering

We filtered irrelevant content using X API's lang field[3], removing non-Arabic tweets and entries with undefined language ("und"). Therefore, tweets containing only mentions, hash-tags, or media links were excluded. Given the limitations of automatic language detection, a sample of tweets was manually verified to ensure data quality.

Phase 2: Data Cleaning

Data cleaning is crucial for reducing noise and enhancing the performance of sentiment classification models. We removed non-Arabic characters, special symbols, numbers, punctuation, extra spaces, and HTML tags. Additionally, retweets (RT) and mentions (@user) were removed.

Phase 3: Handling multilingualism

This study focused on Arabic tweets while retaining embedded Latin-script tokens, as they may provide valuable cues for classification. To manage the multilingual complexities common in Algerian social media, especially code-switching and Arabizi (see Section 2.6.1), we designed a multilingual processing pipeline, detailed in Algorithm 1. The pipeline began by inputting a cleaned tweet and detecting Latin-script segments using the FastText language identification model[4]. Each identified Latin-script segment was then handled based on its type:

- **Translation of Non-Arabic Tweets** Tweets originally written in French or English were translated into Arabic using the deep_translator Python package[5]. This unification under Arabic facilitates more consistent sentiment analysis, though it may occasionally cause some nuance loss (Oueslati et al., 2020).

[3]https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet
[4]https://pypi.org/project/fasttext-langdetect/
[5]https://pypi.org/project/deep-translator/

- **Arabizi Transliteration** As described in Algorithm 1, we normalized Arabizi variants using a custom mapping derived from FastText word embeddings. This step reduced lexical variation by unifying different spellings of the same word, improving dataset consistency. For example, the Arabizi forms m3lich, me3lich, and mehlich all map to the standard Arabic word لابأس ("Never mind"); additional examples are provided in Appendix B. After normalization, Arabizi texts were transliterated into Arabic script using the Yamli API[6].

Finally, all transformed segments were reintegrated into their original context. By explicitly addressing code-switched content, the pipeline ensured consistent handling of multilingual expressions through normalization, language detection, translation, and transliteration. This process preserved valuable lexical information while unifying script and language.

Phase 4: Emoji Processing

Emojis play a significant role in ASA, especially on social media platforms (Alawneh et al., 2024).To preserve their sentiment value, we replaced emojis with descriptive Arabic terms using the Arabic Emoji Sentiment Lexicon (Arab-ESL) (Hakami et al., 2021). For example, the emoji 💔 was mapped to "قلب مجروح", preserving its emotional nuance and improving interpretability.

Recent studies support the emoji replacement approach. In this regard, Khan et al. (2025) reported a 91.85% accuracy boost in a CNN-LSTM model when emoji sentiment was explicitly encoded. Similarly, Alawneh et al. (2024) showed improved performance of Random Forest and BERT-based classifiers on Arabic airline reviews when emojis were integrated into feature extraction.

Phase 5: Stopword Removal

Stopword removal reduces dimensionality by eliminating high-frequency words with low semantic content (Kumar et al., 2025), such as ال, من, هو, و. However, standard lists often overlook crucial negation terms (e.g., لا, لما، لن, ليس, ما، لم, إن, غير) that are vital for sentiment interpretation. Prior studies emphasize customizing stopword lists for morphologically rich languages. For instance, a tailored list improved Bengali sentiment classification (Mostafa et al., 2025), while Saif et al. (2014) showed that generic stoplists could hinder performance and recommended frequency-based filtering for better accuracy. Based on these insights, we implemented a hybrid stopword removal strategy that combined linguistic rules and statistical filtering:

- A manually curated list of Algerian dialect stopwords (e.g., نتا, راك, واش, كلش).

[6]https://www.yamli.com/api/setup/

- A comprehensive MSA stopword list from an open-source repository[7].
- Frequency-based filtering to refine and validate the list.

In total, **804** MSA and Algerian dialect stopwords were removed during preprocessing.

Algorithm 1 Multilingual processing of Algerian dialect tweets. Cleaned texts are iteratively analyzed to detect and translate Latin-script sequences using FastText and Google Translate, with fallback to Yamli correction for Arabizi.

Require: Cleaned tweet t, Arabizi map $\mathcal{M}$ ▷ $\mathcal{M}$: variants → Arabic
Ensure: Processed tweet $\tilde{t}$

$t \leftarrow \textsc{NormaliseVariants}(t, \mathcal{M})$ ▷ Replace every listed variant by its Arabic key
$\mathcal{W} \leftarrow \mathit{split}(t)$ ▷ Tokenise
$\tilde{\mathcal{W}} \leftarrow [\,]$
$\mathcal{B} \leftarrow [\,]$ ▷ Buffer for consecutive Latin words
for all $w \in \mathcal{W}$ **do**
  if w is Arabic **or** w is emoji **then**
    FlushBuffer
    append w to $\tilde{\mathcal{W}}$
  else if w is Latin **then**
    append w to $\mathcal{B}$
  else
    FlushBuffer
    append w to $\tilde{\mathcal{W}}$
  end if
end for
FlushBuffer ▷ Process any remaining Latin buffer
function FlushBuffer
  if $\mathcal{B} \neq \varnothing$ **then**
    $\phi \leftarrow \mathit{join}(\mathcal{B})$ ▷ Concatenate buffered Latin words
    $l \leftarrow \textsc{DetectLanguage}(\phi)$ ▷ FastText Identification
    if $l \in \{en, fr\}$ **then**
      $\tau \leftarrow \textsc{Translate}(\phi)$ ▷ Google Translate
      $\tau \leftarrow \textsc{RemoveNewlines}(\tau)$
      $\tau \leftarrow \textsc{YamliCorrect}(\tau)$ ▷ Catch residual Arabizi
    else
      $\tau \leftarrow \textsc{YamliCorrect}(\phi)$ ▷ Full Arabizi phrase
    end if
    $\tilde{\mathcal{W}} \leftarrow \tilde{\mathcal{W}} \cup \mathit{split}(\tau)$
    $\mathcal{B} \leftarrow [\,]$
  end if
end function
$\tilde{t} \leftarrow \mathit{join}(\tilde{\mathcal{W}})$ **return** $\tilde{t}$

[7]https://github.com/mohataher/arabic-stop-words

Phase 6: Text Normalization

Normalization ensures consistency in text representation by standardizing variations common in Arabic social media texts (Almutairi et al., 2024). Using the Camel Tools library (Obeid et al., 2020), we applied the following steps:

- **Repeated letter reduction** Words with excessive elongation were shortened by reducing consecutive repeated letters to a maximum of two (e.g., كثيييييييراً became كثييراً). This preserves legitimate word forms (e.g., قصص).
- **Elongation character removal (Tatweel)** Elongation characters, commonly used to emphasize a point in social media communication (Albalawi et al., 2021), were eliminated (e.g., ســـلام became سلام).
- **Diacritic removal (Tashkeel)** Short vowel marks (ً ٌ ٍ َ ُ ِ ّ ْ) were stripped to standardize orthography across forms.
- **Letter Standardization** All variants of Alif Hamza (أ, آ, إ) were unified to (أ), Ta Marbouta (ة) was converted to Ha (ه), and Alif Maksoura (ى) to Yah (ي). This normalization reduced vocabulary size and improved model efficiency.

Phase 7: Stemming

To reduce morphological variation and support model generalization, we applied light stemming using the Tashaphyne stemmer (Zerrouki, 2024), which retains the core meaning of words while improving classification performance (Almutairi et al., 2024). However, this stemmer does not fully account for dialect-specific features in Algerian Arabic, such as negation. For instance, in the word "متعمرش", the prefix "م" and the suffix "ش" together express negation, an aspect not effectively handled by the stemmer. Table 6.2 provides examples of tweets before and after preprocessing.

6.2.3 Data Representation

In this section, we evaluate several text vectorization methods for representing Algerian Arabic in our cleaned dataset. These include static embeddings, TF-IDF with N-grams, AraVec, FastText, and AlgVec, as well as the contextualized MARBERT model. The objective is to build a robust, dialect-aware feature representation for ADSA.

Traditional Feature-Based Representations

In ASA, frequency-based methods like n-grams and Term Frequency-Inverse Document Frequency (TF-IDF) are commonly used for text representation (Omar et al., 2025). These techniques encode term frequencies and co-occurrences as sparse vectors. However,

Table 6.2 Examples of tweets before and after the preprocessing pipeline

Original Tweet	After Preprocessing	Preprocessing Task
والخير في ماختاره.. الله#Algerie	والخير في ماختاره الله	Punctuation and Hashtag Removal
الله يخليك و يحفظك عبدو ♥	الله يخليك و يحفظك عبدو قلب ازرق	Emoji Replacement
فنان good work	فنان عمل جيد	Code-switching + Translation
Khouya bzzaffff wlh مبدعين	اخي كثيرا والله مبدعين	Code-switching + Arabizi Normalization
Wch yjik taleb تقرعيجة خفيفة ضريفة	وش يجيك طالب تقرعيجة خفيفة ضريفة	Code-switching + Transliteration
لا لا اريد ان اضحك قليلا	لا لا اريد اضحك قليلا	The stopword "أن" was removed, while "لا" was retained to preserve negation meaning
فاااااااااقوووووووووو	فااقوو	Repeated letter reduction
الحمد لله على كل حال	الحمد، لله، على، كل، حال	Tokenization
ماكانش منها هاذي خرافة	ماكانش من هاذ خراف	Light Stemming

they often struggle with dialectal variation, data sparsity, and semantic nuances, which require extensive manual feature engineering. Despite these limitations, TF-IDF with n-grams remains a strong baseline for traditional classifiers such as SVM and NB, owing to its computational efficiency and interpretability (Kumar et al., 2025). In this study, we employed TF-IDF with n-grams as a baseline to benchmark more advanced embedding techniques.

Dense Word Embeddings

Dense word embeddings provide an alternative to sparse vectors by capturing semantic relationships in high-dimensional continuous space. Below, we describe the embedding techniques used in our study:

AraVec is a pre-trained word embedding model for Arabic NLP . It generates distributed word representations that capture semantic and syntactic relationships. AraVec offers multiple Word2Vec-based models trained on diverse Arabic corpora, including tweets. Its third version includes 16 models trained on over 1.1 billion tokens, providing broad coverage for Arabic text processing (Soliman et al., 2017b).

FastText is an extension of Word2Vec that incorporates subword information by generating embeddings for character n-grams. This approach is especially effective for morphologically rich languages like Arabic. We chose FastText for our deep learning models because its subword modeling captures variations crucial for processing informal Arabic dialect text.

AlgVec Embeddings are custom word embeddings trained on our Algerian dialect X corpus using the Skip-gram algorithm. Skip-gram was preferred over CBOW because it better captures word relationships across diverse contexts (Alosaimi et al., 2024a; Elhassan et al., 2023a), which is essential given the informal and varied nature of Algerian Arabic

on X. To ensure high-quality representations, we:

- Represented words as 300-dimensional dense vectors, following established practices in Arabic embeddings (Soliman et al., 2017b).
- Used zero-padding to handle variable-length sequences, ensuring consistent input sizes for deep learning models.

Contextualized Representations with Transformer Models

We employed MARBERT, a Transformer-based model designed for Arabic NLP (Abdul-Mageed et al., 2021b). It uses self-attention layers to generate context-aware word embeddings. Pre-trained on large Arabic social media corpora, MARBERT effectively handles the dialectal and informal language typical of X posts.

6.2.4 Sentiment Classification Models

We evaluated three main modeling approaches, all trained and tested on our annotated dataset (see Section 6.2.1). Each model followed the preprocessing pipeline described in Section 6.2.2, as illustrated in Figure 6.2. The architectures are summarized below, while their training configurations, parameter settings, and dataset split are detailed in Section 6.3 to ensure full reproducibility.

Classical Machine Learning-based Models

To classify tweets as positive, negative, or neutral, we used two classical ML algorithms: SVM and Multinomial Naive Bayes (MNB), implemented with scikit-learn's default settings.

Deep Learning-based Models

We employed two DL architectures: CNN and Bidirectional LSTM. These models were chosen for their effectiveness in capturing local patterns and sequential dependencies in text (Kumar et al., 2025).

CNNs extract local features through convolutional layers, followed by pooling layers to reduce complexity. We adopted the 18-layer sequential CNN architecture from (Elhassan et al., 2023a), starting with an embedding layer to convert text into dense vectors. The architecture was followed by two 1D convolutional layers with ReLU activation and batch normalization. Max-pooling and dropout layers were applied to prevent overfitting. The model concluded with two fully connected layers and a Softmax output layer for classifying tweets as positive, negative, or neutral.

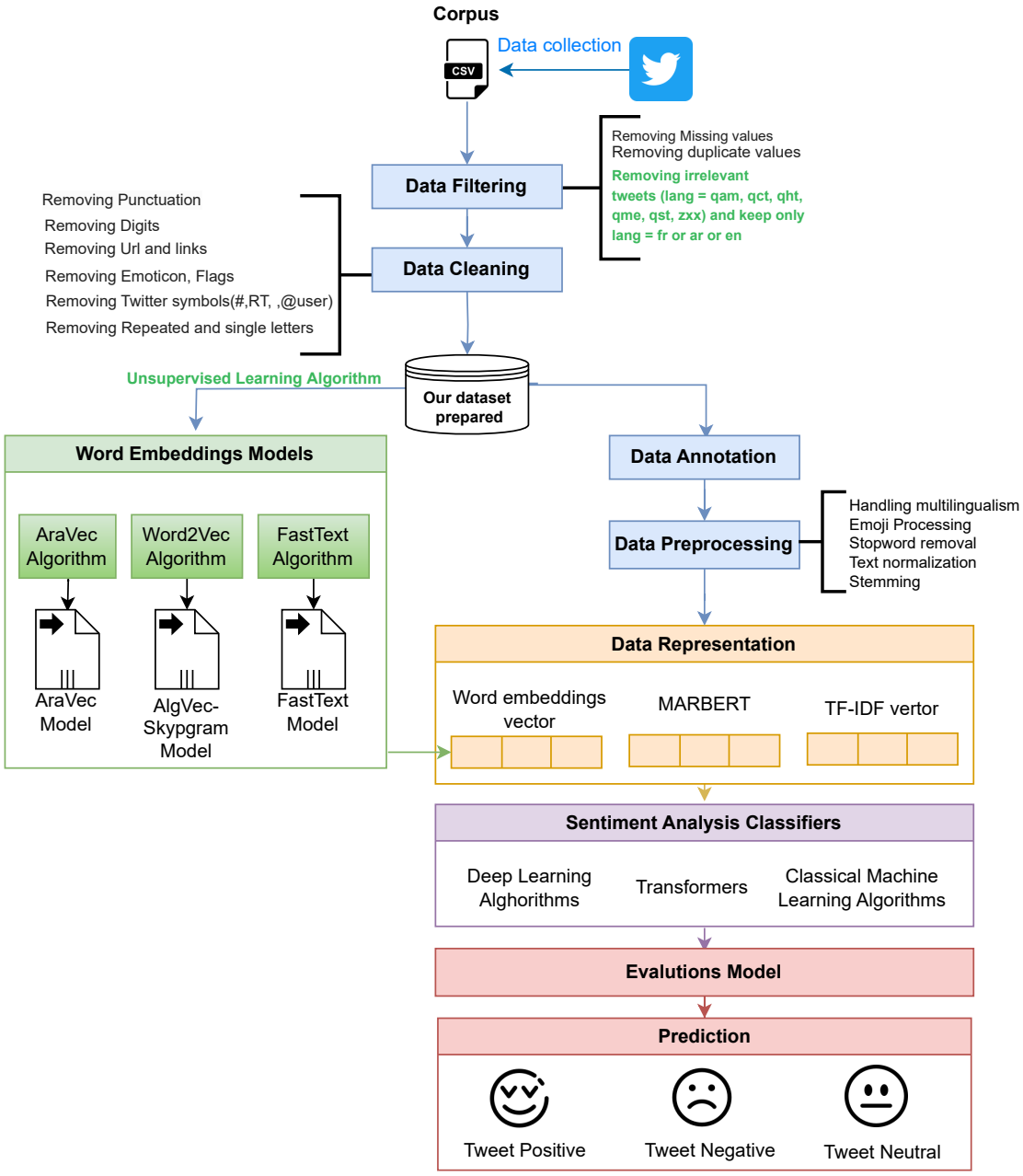


Fig. 6.2 Flow diagram of the proposed architecture, illustrating the main components and data flow for Algerian sentiment analysis on X.

LSTM addresses the vanishing gradient problem through memory cells that capture long-term dependencies in text (Elhassan et al., 2023a). We implemented a BiLSTM architecture, beginning with an embedding layer that incorporated pre-trained word embeddings. A BiLSTM layer with 128 units followed, capturing contextual information from both directions. To reduce overfitting, we included two dropout layers with a rate of 0.3: one after the BiLSTM and another after a dense layer with 64 units. The network

ended with a dense layer using ReLU activation, followed by a Softmax output layer that classified each tweet as positive, negative, or neutral.

Transformer-based Combined Model

To leverage the strengths of Transformer models and sequence learning, we developed a hybrid architecture combining MARBERT with an LSTM layer. This approach is effective for sentiment analysis of low-resource languages like the Algerian dialect (Alosaimi et al., 2024a). The model received tokenized inputs, including input IDs and attention masks, which MARBERT processes to generate rich contextual embeddings. These embeddings passed through a 128-unit LSTM layer with ReLU activation to capture additional sequential information. To prevent overfitting, a dropout layer with a 0.2 rate followed the LSTM. Next, a fully connected layer with 150 neurons and ReLU activation further refined the features, followed by another dropout layer. Finally, a Softmax output layer classified each tweet as positive, negative, or neutral.

Overall, these three modeling approaches provide complementary strengths for sentiment analysis of Algerian Arabic text. This diverse strategy enables benchmarking traditional methods against advanced architectures and evaluating the impact of contextualized embeddings on classification performance.

6.3 Experiments and Results

Building on the model architectures described in Section 6.2.4, this section presents the experimental setup, including the dataset preparation, implementation details, and parameter settings, aimed at ensuring clarity and reproducibility. The experiments were conducted using Google Colab. For classical ML models, we implemented SVM and MNB using the scikit-learn library, while for DL tasks we employed Gensim for word embeddings and Keras (built on top of TensorFlow) for model development.

The experiments used the X4ADSA dataset, which contains 18,589 tweets classified into three sentiment categories: positive (34.31%), negative (33.20%), and neutral (34.31%). The dataset was divided into a training set (90%) and a testing set (10%) to ensure sufficient data for model learning and evaluation.

6.3.1 Model Evaluation

To evaluate the performance of the algorithms used for AD sentiment classification, we used four standard metrics. Accuracy (Eq. 6.1) measures the proportion of correctly classified tweets. Precision (Eq. 6.2) reflects the percentage of relevant results among all positive predictions, indicating the rate of false positives. Recall (Eq. 6.3) measures the ability to identify actual positive samples, capturing the rate of false negatives. Lastly,

F1-score (Eq. 6.4) is the harmonic mean of precision and recall, offering a balanced metric, especially useful for imbalanced datasets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{6.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{6.3}$$

$$F1\text{-score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{6.4}$$

6.3.2 First Experiment: Evaluating Preprocessing and Feature Extraction

Inspired by Albalawi et al. (2021), this experiment investigated the impact of preprocessing techniques (Section 6.2.2) on the performance of ADSA classifiers. We employed two baseline models, SVM and MNB, and evaluated TF-IDF-based feature extraction using unigram, bigram, and trigram representations. As a baseline, both models were first trained on unprocessed data. We then applied TF-IDF to construct sparse feature matrices and systematically examined the effect of individual and combined preprocessing techniques to determine the optimal configuration for sentiment classification.

Results

The MNB model consistently outperformed SVM across all feature extraction techniques, as presented in Table 6.3. The highest accuracy of 73.3% was achieved by MNB using TF-IDF with unigrams. For SVM, accuracy slightly declined with larger n-gram sizes, 72.6% with unigrams, 72.3% with bigrams, and 72.2% with trigrams. MNB followed a similar trend with slightly smaller drops, achieving 73.3%, 73.2%, and 72.9% for unigrams, bigrams and trigrams, respectively.

Table 6.3 Baseline results for SVM and MBN with different feature configurations

Model	Technique	Accuracy (%)
SVM	TF-IDF + Unigram (defaults)	72.6
SVM	TF-IDF + Bigram	72.3
SVM	TF-IDF + Trigram	72.2
MNB	TF-IDF + Unigram (defaults)	**73.3**
MNB	TF-IDF + Bigram	73.2
MNB	TF-IDF + Trigram	72.9

Regarding preprocessing, MNB outperformed SVM even on unprocessed data (76.3% vs. 75.5%). Subsequently, applying individual preprocessing steps resulted in modest gains. Specifically, text cleaning improved accuracy to 76.2% for SVM and 76.4% for MNB. In addition, emoji replacement provided slight improvements, yielding 76.1% for SVM and 76.2% for MNB. Moreover, stopword removal offered a notable boost for MNB, increasing its accuracy to 76.7%. On the other hand, normalization slightly enhanced SVM's performance to 76.5%, while notably, light stemming produced SVM's best individual result at 76.8%. Finally, when all preprocessing techniques were combined, both models achieved their highest performance: 77.2% for SVM and 78.1% for MNB, as illustrated in Figure 6.3.

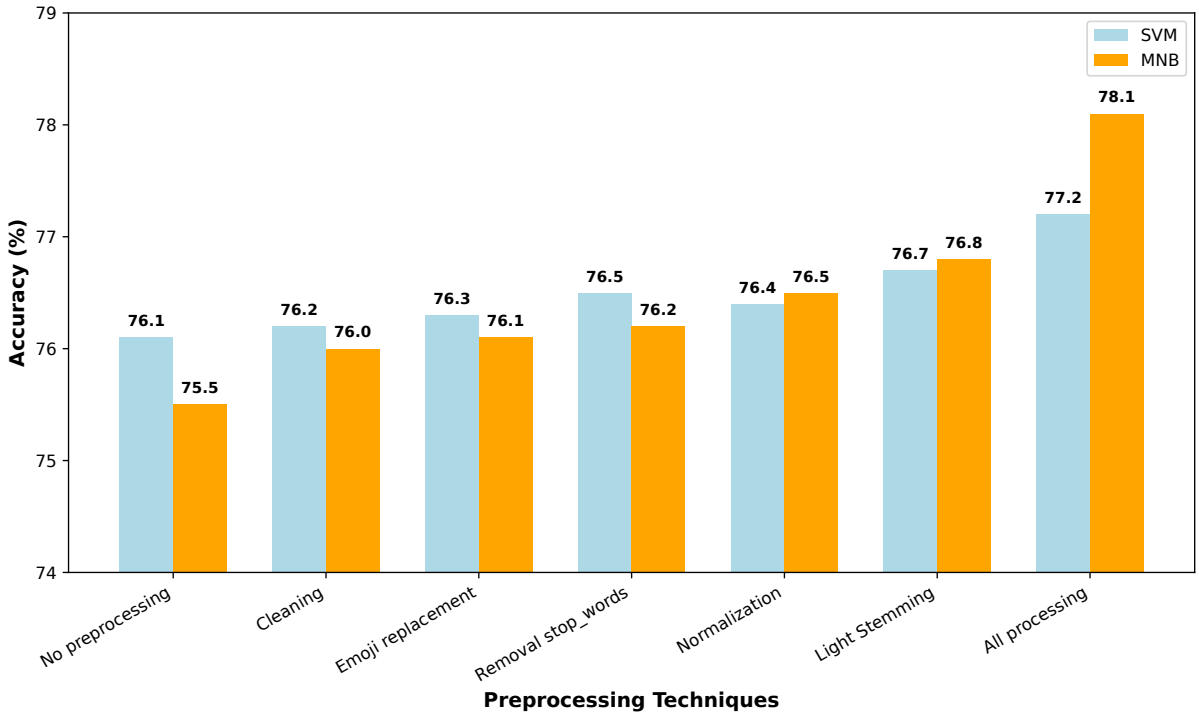


Fig. 6.3 Impact of Preprocessing Techniques on the Performance of SVM and MNB Models

To contextualize our results, we compared the performance of our best model, MNB with optimized preprocessing, to the top model reported by Albalawi et al. (2021), which inspired our work. Although their model achieved higher overall scores, our approach remains competitive, as shown in Table 6.4. It is important to highlight that their study focused on MSA within the health communication domain, utilizing more structured data. In contrast, our work addressed the informal and linguistically diverse nature of the Algerian dialect. These differences in language variety and data characteristics likely account for the observed performance gap.

Table 6.4 Performance comparison between our best model and the best model from Albalawi et al. (2021)

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Our best Model[a]	80.44	80.03	80.24	78.10
The best model of Albalawi et al. (2021)	87.5	84.6	86.0	91.6

6.3.3 Second Experiment: Comparing Word Embedding Models for DL Architectures

In this experiment, we compared the performance of three word embedding models, FastText, AraVec, and our custom AlgVec-Skip-gram, when used as input layers for CNN and LSTM models. These DL architectures were benchmarked against the best-performing traditional classifier from the first experiment.

- The BiLSTM model was trained with a batch size of 128, a dropout rate of 0.3, and for 25 epochs. To ensure a fair comparison, we varied only one hyperparameter at a time while keeping all others fixed.
- The final CNN architecture consisted of convolutional layers with 200 filters and kernel sizes of [3, 4, 5], followed by a fully connected layer with 30 units. To mitigate overfitting, a dropout rate of 0.5 was applied. The CNN model was trained for 10 epochs with a learning rate of 0.001 and a batch size of 50, as detailed in Table 6.5.

Table 6.5 BiLSTM and CNN Model Hyperparameters

Hyperparameter	LSTM Values	CNN Values
Dropout Rate	0.5	0.5
Learning Rate	0.001	0.001
Optimization Algorithm	Adam	Adam
Batch Size	128	50
Number of Epochs	25	25 (EarlyStopping at 10 Epochs)

Results

Among the evaluated embeddings, AraVec combined with BiLSTM achieved the highest performance, with a precision of 89.03%, recall of 90.12%, F1-score of 89.42%, and accuracy of 88%, highlighting its strong ability to represent dialectal Arabic. FastText performed competitively in the CNN setup, with balanced precision and F1-scores around 87.5%, indicating solid general performance. AlgVec, while slightly behind in the CNN setting, delivered robust results with BiLSTM, achieving an F1-score of 87.10%, showcasing its potential as a dialect-specific embedding. Overall, AraVec with BiLSTM emerges as the most effective model configuration, as detailed in Table 6.6.

Table 6.6 Comparison of CNN and BiLSTM models across different embeddings techniques

Model	Embedding	Precision (%)	Recall (%)	F_1score (%)	Accuracy (%)
CNN	AraVec	88.90	87.50	88.00	87.50
	FastText	87.53	87.77	87.36	87.77
	AlgVec	85.30	84.90	85.00	84.85
BiLSTM	AraVec	89.03	90.12	89.42	88.00
	FastText	88.32	87.34	87.80	87.50
	AlgVec	87.75	86.90	87.10	86.80

6.3.4 Third Experiment: MARBERT with LSTM Hybrid Model

Building on the work of Alosaimi et al. (2024a), who employed the AraBERT architecture for sentiment classification, we selected MARBERT due to its extensive pretraining on diverse Arabic dialectal content from social media, offering richer representations for dialect-specific language. For training, we applied:

- Adam optimizer with a learning rate of 2e-5 to ensure stable and effective fine-tuning of MARBERT's pretrained weights.
- Batch size of 32 to balance GPU memory constraints with reliable gradient estimation.
- Input sequences limited to 128 tokens to cover most tweets while reducing padding overhead.
- Class weights integrated into the loss function to address dataset imbalance.
- Frozen MARBERT weights during training to focus learning on the LSTM and classifier layers, minimizing overfitting risks given the moderate dataset size.

Results

The MARBERT with LSTM model demonstrated strong performance on our dataset. It achieved 88.70% precision, 90.25% recall, 89.47% F1-score, and 91.23% accuracy, reflecting a well-balanced capability between precision and recall.

While the F1-scores of the BiLSTM and MARBERT-LSTM models are similar, the MARBERT-based model achieved higher accuracy and recall, as shown in Table Table 6.7. This demonstrates its stronger ability to correctly identify positive and negative sentiments. It also confirms that combining MARBERT's pretrained dialectal embeddings with LSTM's sequential learning yields a robust and superior approach for Algerian Arabic sentiment analysis.

Table 6.7 Comparative Performance Metrics of all models

Model	Feature technique	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
MNB	TF-IDF + Unigram	80.44	80.03	80.24	78.10
BiLSTM	AraVec	89.03	90.12	89.42	88
MARBERT-LSTM	MARBERT	88.70	90.25	89.47	91.23

6.3.5 Cross-Domain Evaluation

To assess the generalizability of our best model, MARBERT-LSTM, we evaluated it on external datasets from different platforms. This evaluation is crucial because language use varies across social media, posing challenges for sentiment models (Alahmadi et al., 2025). To this end, we tested the model on two datasets:

- **YouTube Dataset** A public dataset of 45,000 comments from over 30 Algerian press channels (Benmounah et al., 2023b), labeled as positive (35.5%), negative (42.1%), and neutral (22.4%). The dataset is imbalanced, with negative comments most frequent.
- **X Dataset** A collection of 9,000 tweets (Moudjari et al., 2020) labeled positive (48.3%), negative (29.1%), and neutral (22.6%). This dataset is moderately imbalanced, with positive tweets predominating.

We applied the same preprocessing pipeline and data formatting used during training and split each dataset into 90% for testing and 10% for evaluation, consistent with our main experimental setup.

Results

The model demonstrated strong transferability on the YouTube dataset, achieving an F1-score of 79.70% and an accuracy of 82.15%. However, results dropped on the X dataset, with an F1-score of 64.57% and accuracy of 71.90%, mainly due to class imbalance. The neutral class was underrepresented, with only 257 samples left after removing noisy tweets (e.g., URLs, video links). This led to poor recognition of neutral sentiment, with an F1-score of just 5%.

To address this, we applied the Synthetic Minority Oversampling Technique (SMOTE) on MARBERT embeddings in the training set to generate synthetic neutral samples and balance the dataset. This intervention significantly improved performance on the X dataset, raising the F1-score to 75.49% and accuracy to 76.72%, thereby closing the gap with results obtained on our dataset and the YouTube data (see Table 6.8).

Table 6.8 Performance of MARBERT-LSTM on External Datasets, Highlighting SMOTE Oversampling Effects

Dataset	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Own Dataset	88.70	90.25	89.47	91.23
YouTube Dataset	79.50	80.00	79.70	82.15
X Dataset[a]	66.83	62.47	64.57	71.90
X Dataset[b]	74.28	76.72	75.49	76.65

6.4 Discussion

This chapter comprehensively evaluated different modeling strategies for Algerian Dialect Sentiment Analysis, spanning traditional machine learning, word embedding approaches, and transformer-based architectures.

Our findings confirm the importance of thorough preprocessing and feature engineering in traditional classifiers, with the MNB model benefiting most from combined preprocessing steps and unigram TF-IDF features. This underscores the continued relevance of language-specific normalization techniques, such as light stemming and letter normalization, for Arabic dialects.

When comparing word embeddings, our experiments showed that domain-specific and dialect-aware embeddings, such as AlgVec, enhance the performance of deep learning models. However, the MARBERT-LSTM hybrid model achieved particularly strong and balanced results, effectively capturing subtle sentiment distinctions, as reflected in its high recall and accuracy. These findings highlight the advantage of combining pretrained transformer embeddings with sequence modeling to address the complexity and variability of low-resource dialectal data.

Overall, our results demonstrate that while classical and embedding-based models remain valuable, transformer-based models pretrained on dialectal corpora offer a significant advantage. These findings highlight their potential to enhance real-world sentiment analysis applications for Algerian Arabic.

6.4.1 Comparison with prior studies

We compared our MARBERT-LSTM model with recent transformer-based approaches for sentiment analysis in Arabic dialects. Our model attains 91.23% accuracy and an F1-score of 89.47%, outperforming comparable models, as shown in Table 6.9. This success stems from combining MARBERT's pretrained dialectal embeddings with an LSTM decoder.

For instance, Benmounah et al. (2023b) obtained 81.74% accuracy on Algerian YouTube comments using BERT-based models, over 10 points lower than our result. Similarly, Matrane et al. (2024) reached 90.45% accuracy on Moroccan dialect tweets, highlighting the robustness of our approach given the linguistic proximity of Moroccan and Algerian di-

alects.

While Sherif and Sabty (2024) attained a higher F1-score of 92.54% on a hybrid Egyptian-English dataset, their results benefit from English content, whereas our model operates solely on dialectal Arabic. Additionally, Omar et al. (2025) achieved maximum accuracies of 90.40% using pretrained models like MARBERT and CAMeLBERT on various Arabic dialects. Their results remain below ours.

Although Alosaimi et al. (2024a) reached 93.76% accuracy using AraBERT on a large multi-domain Saudi corpus, their dataset does not capture challenges such as Arabizi and emoji use, which are common in our Algerian dataset.

Overall, these comparisons confirm that transformer-based models excel in Arabic sentiment analysis by effectively handling informal, noisy dialectal expressions typical of Algerian Arabic.

Table 6.9 Comparative performance analysis of the proposed model and previous studies

Reference	Language	Model	DataSet	Embedding	Acc[a]
(Benmounah et al., 2023b)	Algerian	BERT-based	YouTube	–	81.74
(Alosaimi et al., 2024a)	Saudi Arabian	AraBERT-LSTM	SS2030 AHS ASTC	CBOW Skipgram AraBERT	93.76
(Matrane et al., 2024)	Moroccan	QaRIB DarijaBERT AraBERT	FB MSAC MAC MSTD	AraBERT DarijaBERT QaRIB	90.45
(Sherif and Sabty, 2024)	Egyptian/English	BiLSTM-Attention Hybrid-Transformer	EESA	FastText AraBERT	92.54[b]
(Omar et al., 2025)	MSA/Dialects	CAMeLBERT XLM-R MARBERT	ASTD ASAD TEAD	AraBERT	90.40
Our Study	Algerian	MARBERT-LSTM	X4ADSA	MARBERT	91.23

6.4.2 Cross-Domain Performance Discussion

Testing our MARBERT-LSTM model on the external YouTube dataset showed a drop in F1-score from 89.47% to 79.70%, mainly due to two factors:

- YouTube comments are often much longer than tweets; for example, some misclassified comments exceed 700 characters. Since MARBERT has a fixed maximum input length (approximately 512 tokens) and truncates longer texts. As a result, important contextual information is sometimes lost, which reduces classification accuracy.
- YouTube comments tend to contain more nuanced or mixed sentiments within a single comment, while tweets are generally shorter and more polarized.

These observations align with Shevtsov et al. (2023), who showed that although topics overlap across platforms, user interaction and sentiment expression differ significantly. The longer, varied linguistic style on YouTube challenges models trained solely on tweet data.

In contrast, the evaluation of the model on the external X dataset revealed a notable limitation in handling severe class imbalance, particularly due to the underrepresentation of the neutral sentiment class. This imbalance led to a marked drop in F1-score despite the robust MARBERT-LSTM architecture and domain-specific embeddings. Applying SMOTE improved performance, as illustrated in Figure 6.4, showing the model's sensitivity to class imbalance and the importance of targeted augmentation.

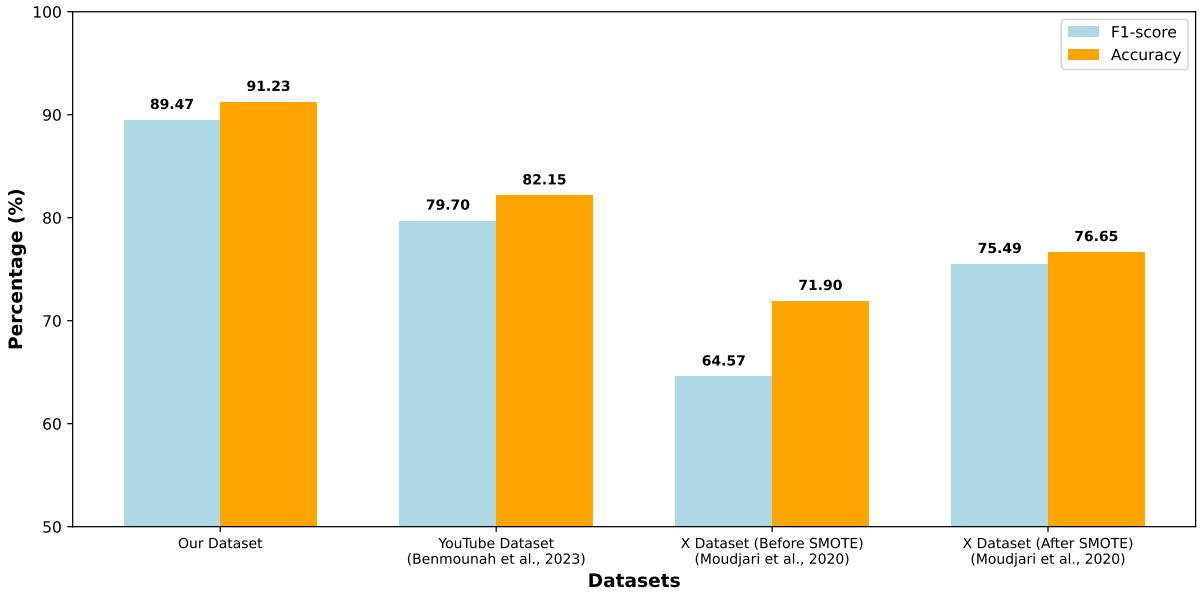


Fig. 6.4 Performance of MARBERT-LSTM on Cross-Domain Sentiment Classification and Improvement after SMOTE Oversampling

6.4.3 Error Analysis and Model Limitations

Despite the strong overall performance of the MARBERT-LSTM model, frequent misclassifications highlight its moderate limitations in handling informal Algerian Arabic. The main issues observed are summarized in Table 6.10 and detailed below:

- Sarcasm and irony: The model struggled to capture implied or context-dependent meanings. For instance, the phrase "الجزاير الجديدة" ("The New Algeria"), often used ironically in political contexts, was incorrectly classified as positive.
- Metaphorical expressions: Figurative language was often misinterpreted. For example, "برنامج قمة" ("Top program") was sometimes labeled as neutral, as the model interpreted "قمة" literally rather than recognizing its figurative or emotionally charged meaning.

- Dialect-specific semantics: Certain words differ in meaning between Algerian Arabic and MSA. For instance, in the phrase "غير كمل ميرسي" ("Just continue, thanks"), the word "غير" expresses encouragement in the dialect, but its negative sense in MSA led to confusion. The model frequently failed to disambiguate these conflicting cues.
- Preprocessing artifacts: Some errors stemmed from preprocessing. Removing punctuation, such as question marks, altered the intended meaning of tweets. For example, the tweet "موقف الجزائر من تصريحات ماكرون؟" ("What is Algeria's stance on Macron's statements?") lost its interrogative form after punctuation removal, leading the model to misclassify it as negative. Similarly, replacing emojis with literal words introduced ambiguity; for example, substituting the fire emoji 🔥 with "نار" obscured its intended positive or intensifying sentiment.

These observations highlight that, despite strong overall performance, the model has difficulty handling figurative language, punctuation, and dialect-specific cues. Addressing these limitations requires more robust preprocessing and models tailored to the linguistic complexity of Algerian Arabic, which we leave for future work.

Table 6.10 Examples of Misclassified Tweets for Error Analysis

Tweet	True Label	Predicted Label
القراية الزينة لبقات في الجزاير The good education that remained in Algeria	Negative	Neutral
الجزائر لا تقع..الجزائر في القمة مع ليبيا والسعودية ومصر Algeria does not fall... it is at the top with Libya, KSA, Egypt	Positive	Neutral
حشاك ياهذا You are above this	Positive	Neutral
200k بصحتك 🔥 Congrats on your 200k	Positive	Neutral
غير كمل ميرسي Just continue, thanks	Positive	Neutral
موقف الجزائر من تصريحات ماكرون؟ What is Algeria's stance on Macron's statements?	Neutral	Negative

6.5 Conclusion

This study advances sentiment analysis for low-resource and informal dialects by introducing X4ADSA, a large annotated corpus of 18,589 Algerian Arabic tweets. It also presents a specialized preprocessing pipeline to tackle challenges like Arabizi, code-switching, and dialectal variation. Our comprehensive approach to multilingualism, through transliteration, translation, and a custom Arabizi normalization, substantially enhances the model's understanding of transliterated dialect words.

The proposed MARBERT-LSTM model outperforms recent state-of-the-art approaches on both Arabic and dialectal datasets, highlighting the importance of dialect-specific adaptation and advanced sequence modeling. Beyond sentiment analysis, this work establishes crucial resources and methodologies that can accelerate NLP development for other under-resourced dialects and languages.

Future work will focus on enhancing the detection of subtle dialectal nuances and improving generalization across different platforms and longer texts. We plan to expand the training dataset to achieve better Arabizi normalization and investigate adaptive strategies for varying input lengths. Moreover, we aim to extend this approach to other Maghrebi dialects and develop complementary NLP tools, including text normalization and machine translation, to better support Algerian and related dialects. Finally, improving human-machine interfaces will be prioritized to increase the usability and real-world impact of these tools for low-resource communities.

Chapter 7

Integrating Sentiment Analysis into HMI

7.1 Introduction

This chapter investigates the integration of Algerian Arabic sentiment analysis into a Human-Machine Interaction framework, moving beyond standalone sentiment classification toward practical, user-centered system behavior. Unlike traditional interfaces that rely exclusively on explicit commands, the proposed approach treats sentiment as an implicit interaction signal, enabling systems to adapt their responses dynamically according to the user's affective state.

In human-centered computing, affect-aware interaction has been shown to improve usability, reduce cognitive load, and enhance user satisfaction. However, most existing sentiment-aware HMI systems are designed for high-resource languages and standardized linguistic input. In the Algerian context, user interaction is characterized by informal language, dialectal variation, frequent code-switching, and the widespread use of Arabizi. These properties significantly limit the applicability of generic sentiment models in real-world interactive systems.

Building on the sentiment analysis models, embeddings, and preprocessing strategies developed in the preceding chapters, this chapter demonstrates how dialect-aware sentiment analysis can be operationalized within an interactive system. By supporting both Arabic script and Arabizi, the proposed framework aligns with the natural communication practices of Algerian users and enables robust interpretation of subjective textual input in real time.

The objective of this chapter is not to introduce new sentiment models, but to show how previously developed resources, particularly AlgVec embeddings and fine-tuned transformer-based classifiers, can be integrated into an adaptive HMI pipeline. Through a functional prototype, we illustrate how sentiment predictions can directly influence interaction strategies, thereby positioning sentiment analysis as a core component of intelligent, user-adaptive interfaces rather than a purely analytical task.

7.2 The Role of Sentiment as an Interaction Signal

While the theoretical foundations of human–machine interaction and affective computing were introduced in Section 2.7, this section focuses on the practical use of subjectivity analysis within an interactive system. In this context, sentiment is not treated as a post-hoc analytical metric, but rather as a real-time input that directly informs system behavior.

In sentiment-aware HMI, textual user input provides indirect yet valuable feedback regarding affective states such as satisfaction, neutrality, or frustration. These signals can be exploited to refine system responses, prioritize actions, or dynamically adjust the interaction flow, thereby aligning system outputs with the user's emotional state (Ghosh, 2023).

The effectiveness of such interaction mechanisms is highly dependent on linguistic accuracy. In the Algerian context, the widespread use of dialectal expressions, code-switching, and Arabizi significantly limits the applicability of generic sentiment models. As a result, integrating a localized, dialect-aware sentiment model is essential to correctly capture users' nuanced subjective expressions.

This perspective differentiates the proposed system from conventional sentiment analysis pipelines by emphasizing real-time interpretability and actionable affective feedback within interactive environments, rather than offline sentiment evaluation alone.

7.3 System Architecture and Integration

This section details the methodology for incorporating Algerian Dialect Sentiment Analysis into an interactive framework, organized into preprocessing, modeling, and decision-making modules.

7.3.1 Linguistic Preprocessing for Dialectal Robustness

To ensure reliability, raw user input undergoes a specialized pipeline designed for Algerian informal speech. As detailed in Section 6.2.2, this includes:

1. Normalization: Standardizing orthographic variations in both Arabic script and Arabizi.

2. Tokenization and Emoji Handling: Extracting semantic meaning from non-textual symbols which carry high emotional weight in Algerian social media.

3. Transliteration: Converting Arabizi into a unified representation for the model.

7.3.2 Sentiment Modeling

The sentiment prediction module builds directly on the resources developed earlier in this thesis. Specifically, it combines the **AlgVec** word embeddings introduced in Chapter 4 with a transformer-based architecture based on MARBERT, fine-tuned on the Algerian Dialect Sentiment Analysis dataset.

AlgVec provides dialect-aware lexical representations for both Arabic script and Arabizi, while MARBERT contributes contextualized embeddings learned from large-scale Arabic social media data. This hybrid design allows the system to capture fine-grained semantic nuances, informal expressions, and code-switched patterns characteristic of Algerian user input.

The model outputs a probability distribution over three sentiment classes (positive, negative, and neutral). Rather than serving as an end result, these probabilities constitute the quantitative input for the decision layer, where sentiment information is transformed into adaptive interaction strategies.

7.4 Decision Layer for Adaptive Interaction

The core contribution of this interface is the translation of sentiment probabilities into actionable system behaviors. This bridge ensures that the machine's response is contextually appropriate and subjectivity-aware.

7.4.1 Sentiment to Action Mapping

The system employs a rule-based mapping logic to determine the interaction strategy. Rather than producing static replies, system responses are adapted according to the detected sentiment polarity, as summarized in Table 7.1.

Table 7.1 Mapping sentiment predictions to interaction strategies.

Detected Sentiment	Interaction Strategy	System Action Example
Positive	Reinforcement	Affirmative tone; encouraging further input
Negative	Support / Recovery	Empathetic tone; offering guidance or clarification
Neutral	Informative	Concise, task-oriented standard responses

Figure 7.1 presents the implemented sentiment-aware HMI prototype, illustrating how sentiment predictions are displayed and used to generate adaptive responses in real time.

7.4.2 Contextual Response Adaptation

Beyond simple sentiment-to-action mapping, adaptive response mechanisms incorporate contextual information and interaction history. Responses can be adjusted based on

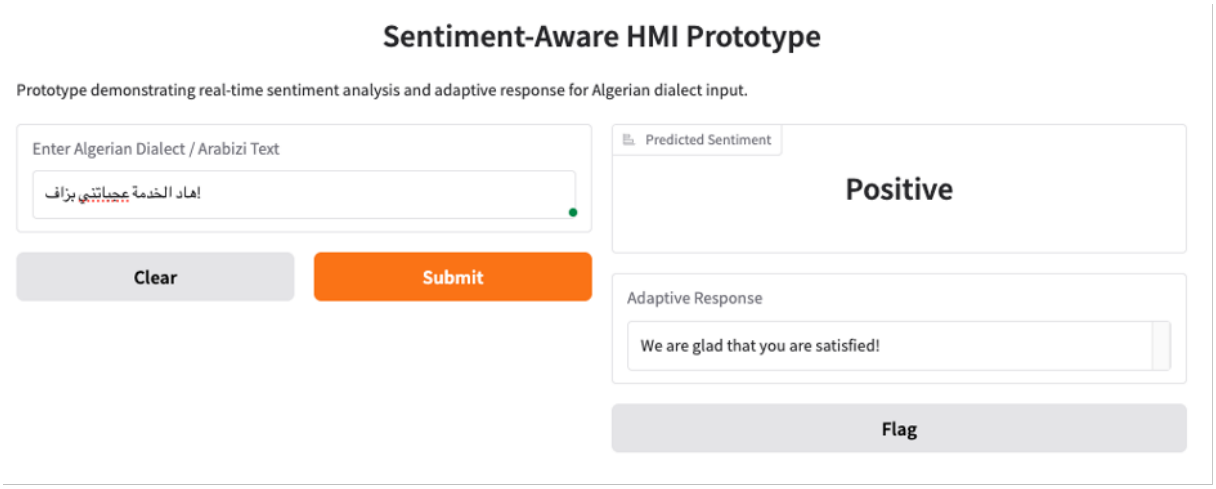


Fig. 7.1 The proposed sentiment-aware HMI prototype for Algerian dialect input. The system displays the predicted sentiment and adaptive response in real time.

previous user inputs, enabling coherent multi-turn interactions. Probabilistic sentiment outputs may also modulate response intensity, allowing nuanced reactions in cases of ambiguous or mixed sentiment.

7.5 Conclusion

This chapter demonstrated how sentiment analysis for the Algerian dialect can be operationalized as a core component of human-machine interaction, rather than treated as an isolated classification task. By integrating dialect-aware preprocessing, AlgVec embeddings, and a transformer-based sentiment model within an adaptive decision layer, the proposed framework enables interactive systems to respond dynamically to users' affective states.

The presented prototype illustrates the practical feasibility of deploying sentiment-aware interaction in low-resource, linguistically diverse settings. Importantly, this chapter builds directly on the methodological contributions of the thesis, showing how sentiment models, embeddings, and preprocessing strategies can be translated into real-world interactive behavior.

While the current implementation focuses on textual input, the proposed architecture is extensible and provides a foundation for future multimodal HMI systems incorporating speech, visual cues, and contextual signals. Overall, this chapter reinforces the central thesis that dialect-aware sentiment analysis plays a critical role in the development of intelligent, user-centered systems for underrepresented language varieties.

Chapter 8

Conclusion and Future Works

This thesis investigated the challenges and opportunities of sentiment and subjectivity analysis in the Algerian Arabic dialect, aiming to enhance human-machine interaction for speakers of this dialect. Key challenges addressed include non-standard orthography, dialectal variation, and the widespread use of Arabizi in social media texts.

A major contribution is **AlgVec**, a set of word embedding models trained on a large corpus of Algerian dialect text, covering both Arabic and Arabizi. These embeddings were evaluated intrinsically, via word similarity and nearest-neighbor retrieval, and extrinsically through sentiment analysis. Results show that AlgVec effectively captures dialectal and sentiment-specific nuances, providing a robust foundation for downstream NLP tasks.

Another important contribution is the creation of a publicly available dataset of 18,772 Algerian dialect tweets, manually annotated for sentiment (positive, negative, neutral). This dataset, encompassing both Arabic script and Arabizi, offers a valuable resource for future research in dialectal NLP and HMI.

We also proposed a dynamic, attention-based stopword extraction approach, improving text classification performance by identifying contextually significant words. When combined with deep learning architectures, this method enhanced both model accuracy and interpretability.

Furthermore, transfer learning approaches were explored for Arabizi sentiment analysis. The hybrid MARBERT–LSTM model, leveraging AlgVec embeddings, achieved the best performance overall, surpassing classical machine learning methods and existing Arabic embeddings on both in-domain and cross-domain datasets.

Overall, this work advances both linguistic resources and computational methods for the Algerian dialect, enabling more accurate sentiment analysis and facilitating user-centered HMI. The datasets, embeddings, and preprocessing pipelines developed here provide a solid foundation for future studies in dialectal NLP.

Future research directions include:

- Extending models to other Arabic dialects.

- Investigating multilingual transfer learning for low-resource dialects.
- Developing real-time, sentiment-aware human-machine interfaces, potentially incorporating multimodal signals.

In conclusion, combining large-scale linguistic resources, tailored embedding techniques, and state-of-the-art deep learning models significantly improves machines' ability to understand and respond to sentiment in Algerian Arabic, advancing toward more natural and effective human-machine interaction.

Bibliography

Mohamed Abd-Elnabi, Hani Mahdi, and Ahmed Badr. From latin to arabic: Innovative approaches for retrieving original texts and vectorizing franco-arabic. *Indonesian Journal of Computing and Science Engineering*, 14:289–300, 04 2023. doi:10.21817/indjcse/2023/v14i2/231402091.

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. Dziribert: a pre-trained language model for the algerian dialect, 2022.

Amin Abdedaiem, Abdelhalim Hafedh Dahou, Mohamed Amine Cheragui, and Brigitte Mathiak. Fassila: A corpus for algerian dialect fake news detection and sentiment analysis. *Procedia Computer Science*, 244:397–407, 2024a. ISSN 1877-0509. doi:10.1016/j.procs.2024.10.214.

Amin Abdedaiem, Abdelhalim Hafedh Dahou, Mohamed Amine Cheragui, and Brigitte Mathiak. Fassila: A corpus for algerian dialect fake news detection and sentiment analysis. *Procedia Computer Science*, 244:397–407, 2024b. ISSN 1877-0509. doi:10.1016/j.procs.2024.10.214.

Houssem Abdellaoui and Mounir Zrigui. Using tweets and emojis to build tead: an arabic dataset for sentiment analysis. *Computación y Sistemas*, 22(3), September 2018. ISSN 2007-9737. doi:10.13053/cys-22-3-3031.

Adel Abdelli, Fayçal Guerrouf, Okba Tibermacine, and Belkacem Abdelli. Sentiment analysis of arabic algerian dialect using a supervised method. In *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, pages 1–6, Taza, Morocco, 2019. IEEE. doi:10.1109/ISACS48493.2019.9068897.

Muhammad Abdul-Mageed, Shady Elbassuoni, Jad Doughman, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Yorgo Zoughby, Ahmad Shaher, Iskander Gaba, Ahmed Helal, and Mohammed El-Razzaz. Dialex: A benchmark for evaluating multidialectal arabic word embeddings. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 11–20, Kyiv, Ukraine (Virtual), 2021a. Association for Computational Linguistics. doi:10.48550/arXiv.2011.10970.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online, August 2021b. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.551.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. Nadi 2022: The third nuanced arabic dialect identification shared task. In Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, and Wajdi Zaghouani, editors, *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.wanlp-1.9.

Nawaf A. Abdulla, Nizar A. Ahmed, Mohammed A. Shehab, and Mahmoud Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–6, Amman, Jordan, 2013. IEEE. doi:10.1109/AEECT.2013.6716448.

Abdulqader Mohsen Abdullatif Ghallab and Yousef Ali. Arabic sentiment analysis: A systematic literature review. *Applied Computational Intelligence and Soft Computing*, 2020:1–21, 1 2020. ISSN 1687-9724. doi:10.1155/2020/7403128.

Mohamed Elhag M. Abo, Nordiana Ahmad Kharman Shah, Vimala Balakrishnan, Mohamed Kamal, Ahmed Abdelaziz, and Khalid Haruna. Ssa-sda: Subjectivity and sentiment analysis of sudanese dialect arabic. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–5, Sakaka, Saudi Arabia, 2019. IEEE. doi:10.1109/ICCISci.2019.8716466.

Mansour O. Aboelela, E. Talaat, and R. Bustami. Transformer-based ensemble model for dialectal arabic sentiment classification. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2025. doi:10.1145/3638285.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. *LSTM-CNN Deep Learning Model for Sentiment Analysis of Dialectal Arabic*, pages 108–121. Springer, 10 2019. ISBN 978-3-030-32958-7. doi:10.1007/978-3-030-32959-4_8.

Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. Identifying sentiments in algerian code-switched user-generated comments. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2698–2705, Marseille, France, 2020. European Language Resources Association. `https://aclanthology.org/2020.lrec-1.328/`.

Mohamed Ahmed and Youssef Elhadj. Arabic dialect identification in social media: A comparative study of word embedding and deep learning approaches. *TechScience*, 10 (2):123–135, 2024. doi:10.1234/techsci.2024.10.2.123.

Ibrahim Al-Huri. Arabic language: Historic and sociolinguistic characteristics. *English Literature and Language Review.*, 1:28–36, 2015. ISSN 2412-1703. doi:10.13140/RG.2.2.16163.66089/1.

Khaled Alahmadi, Sultan Alharbi, Juan Chen, and Xianzhi Wang. Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. *Social Network Analysis and Mining*, 15(1):45, 2025. ISSN 1869-5469. doi:10.1007/s13278-025-01461-8.

Hussam Alawneh, Ahmad Hasasneh, and Mohammed Maree. On the utilization of emoji encoding and data preprocessing with a combined cnn-lstm framework for arabic sentiment analysis. *Modelling*, 5(4):1469–1489, 2024. ISSN 2673-3951. doi:10.3390/modelling5040076.

Abdulaziz M. Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. *A Combined CNN and LSTM Model for Arabic Sentiment Analysis*, pages 179–191. Springer International Publishing, Cham, 2018. ISBN 9783319997407. doi:10.1007/978-3-319-99740-7_12.

Yahya Albalawi, Jim Buckley, and Nikola S. Nikolov. Investigating the impact of preprocessing techniques and pre-trained word embeddings in detecting arabic health information on social media. *Journal of Big Data*, 8(1):95, December 2021. ISSN 2196-1115. doi:10.1186/s40537-021-00488-w.

Abdullah Aldumaykhi, Saad Otai, and Abdulkareem Alsudais. Comparing open arabic named entity recognition tools, 2022. URL `https://arxiv.org/abs/2205.05857`.

Cagatay Catal Alexander Ligthart and Bedir Tekinerdogan. Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, 54:4997–5053, 10 2021. ISSN 15737462. doi:10.1007/s10462-021-09973-3.

Alaa Alharbi and Mark Lee. Kawarith: an Arabic Twitter corpus for crisis events. In Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors, *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 42–52, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wanlp-1.5/`.

Omar Alharbi. Negation handling in machine learning-based sentiment classification for colloquial arabic. *International Journal of Operations Research and Information Systems*, 11, 01 2020. doi:10.4018/IJORIS.2020100102.

Azzah Allahim and Ahmed Cherif. Advancing arabic word embeddings: A multi-corpora approach with optimized hyperparameters and custom evaluation. *Applied Sciences*, 14 (23):11104, 2024. doi:10.3390/app142311104.

Latifah Almuqren, Arwa Alzammam, Shahad Alotaibi, Alexandra Cristea, and Sarah Alhumoud. A review on corpus annotation for arabic sentiment analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10283 LNCS, pages 215–225, Cham, 2017. Springer Verlag. ISBN 9783319585611. doi:10.1007/978-3-319-58562-8_17.

Tahani Almutairi, Shireen Saifuddin, Reem Alotaibi, Shahendah Sarhan, and Sarah Nassif. Preprocessing techniques for clustering arabic text: Challenges and future directions. *International Journal of Advanced Computer Science and Applications*, 15(8), 2024. doi:10.14569/IJACSA.2024.01508126.

Abdullah Hussein Ali Alnosairee and Ni Wayan Sartini. A sociolinguistics study in arabic dialects. *PRASASTI Journal of Linguistics*, 6(1), 2021. doi:10.20961/prasasti.v6i1.43127. URL `https://creativecommons.org/licenses/by-sa/4.0/`.

Khaled Mohammad Alomari, Hatem M. ElSherif, and Khaled Shaalan. Arabic tweets sentimental analysis using machine learning. In Salem Benferhat, Karim Tabia, and Moonis Ali, editors, *Advances in Artificial Intelligence: From Theory to Practice*, pages 602–610, Cham, 2017. Springer International Publishing. ISBN 978-3-319-60042-0. doi:10.1007/978-3-319-60042-0_66.

Wael Alosaimi, Hager Saleh, Ali A. Hamzah, Nora El-Rashidy, Abdullah Alharb, Ahmed Elaraby, and Sherif Mostafa. Arabbert-lstm: improving arabic sentiment analysis based on transformer model and long short-term memory. *Frontiers in Artificial Intelligence*, 7, 2024a. ISSN 2624-8212. doi:10.3389/frai.2024.1408845.

Wafa Alosaimi, Hatem Saleh, Ahmed A. Hamzah, Noura El-Rashidy, Abdullah Alharb, Ahmed Elaraby, and Samy Mostafa. Arabbert-lstm: improving arabic sentiment analysis based on transformer model and long short-term memory. *Frontiers in Artificial Intelligence*, 7:1408845, 2024b. doi:10.3389/frai.2024.1408845.

Areeb Alowisheq, Nora Al-Twairesh, Mawaheb Altuwaijri, Afnan Almoammar, Alhanouf Alsuwailem, Tarfa Albuhairi, Wejdan Alahaideb, and Sarah Alhumoud. Marsa: Multi-domain arabic resources for sentiment analysis. *IEEE Access*, 9:142718–142728, 2021. ISSN 2169-3536. doi:10.1109/ACCESS.2021.3120746.

Arwa Alqarni and Atta Rahman. Arabic tweets-based sentiment analysis to investigate the impact of covid-19 in ksa: A deep learning approach. *Big Data and Cognitive Computing*, 7(1), 2023. ISSN 2504-2289. doi:10.3390/bdcc7010016.

Mohamed Aly and Amir Atiya. Labr: A large scale arabic book reviews dataset. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. `https://aclanthology.org/P13-2088`.

Djegdjiga Amazouz. *Linguistic and phonetic investigations of French-Algerian Arabic code-switching: Large corpus studies using automatic speech processing.* Phd thesis, Université Sorbonne Nouvelle – Paris 3, 2019. https://www.afcp-parole.org/doc/theses/these_DA19.pdf.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (OSACT4)*, pages 9–15, Marseille, France, 2020. European Language Resource Association. doi:10.48550/arXiv.2003.00104.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding, 2021.

Aamir Anwar, Ikram Ur Rehman, Moustafa M. Nasralla, Sohaib Bin Altaf Khattak, and Nasrullah Khilji. Sentiment analysis and student emotions: Improving satisfaction in online learning platforms. In *2023 IEEE International Smart Cities Conference (ISC2)*, pages 1–7, 2023. doi:10.1109/ISC257844.2023.10293422.

Gaétan Baert, Souhir Gahbiche, Guillaume Gadek, and Alexandre Pauchet. Arabizi language models for sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 592–603, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.51.

Laith H. Baniata and Sangwoo Kang. Switch-transformer sentiment analysis model for arabic dialects that utilizes a mixture of experts mechanism. *Mathematics*, 12(2):242, 2024. ISSN 2227-7390. doi:10.3390/math12020242.

Abubakar Salisu Bashir, Abdulkadir Abubakar Bichi, and Alhassan Adamu. Automatic construction of generic hausa language stop words list using term frequency-inverse document frequency. *Journal of Electrical Systems and Information Technology*, 11(1): 58, 2024. doi:10.1186/s43067-024-00187-5.

Salima Behdenna, Fatiha Barigou, and Ghalem Belalem. Sentiment analysis at document level. In *Smart Trends in Information Technology and Computer Communications*, pages 159–168, Singapore, 08 2016. Springer Nature Singapore. ISBN 978-981-10-3432-9. doi:10.1007/978-981-10-3433-6_20.

Zakaria Benmounah, Abdennour Boulesnane, Abdeladim Fadheli, and Mustapha Khial. Sentiment analysis on algerian dialect with transformers. *Applied Sciences*, 13(20), 2023a. ISSN 2076-3417. doi:10.3390/app132011157.

Zakaria Benmounah, Abdennour Boulesnane, Abdeladim Fadheli, and Mustapha Khial. Sentiment analysis on algerian dialect with transformers. *Appl. Sci.*, 13(20):11157, 2023b. doi:10.3390/app132011157.

Mohamed Benrabah. *Language Conflict in Algeria*. Multilingual Matters, Bristol, Blue Ridge Summit, 2013. ISBN 9781847699657. doi:10.21832/9781847699657.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. Swan and ArabicMTEB: Dialect-aware, arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4654–4670, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi:10.18653/v1/2025.findings-naacl.263.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. Better document-level sentiment analysis from RST discourse parsing. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1263. URL `https://aclanthology.org/D15-1263`.

Thulasi Bikku, Jyothi Jarugula, Lavanya Kongala, Navya Deepthi Tummala, and Naga Vardhani Donthiboina. Exploring the effectiveness of bert for sentiment analysis on large-scale social media data. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–4, 2023. doi:10.1109/CONIT59222.2023.10205600.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi:10.1162/tacl_a_00051.

Venkateswarlu Bonta, Nandhini Kumaresh, and N. Janardhan. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6, March 2019. ISSN 2249-0701. doi:10.51983/ajcst-2019.8.S2.2037.

Yosra Bouassida and Hakim Mezali. Enhancing twitter sentiment analysis using hybrid transformer and sequence models. *Jpn. J. Res.*, 6, January 2025. doi:10.33425/2690-8077.1163.

Soumia Bougrine, Aicha Chorana, Abdallah Lakhdari, and Hadda Cherroun. Toward a web-based speech corpus for algerian dialectal arabic varieties. In Nizar Habash, Mona Diab, Kareem Darwish, Wassim El-Hajj, Hend Al-Khalifa, Houda Bouamor, Nadi Tomeh, Mahmoud El-Haj, and Wajdi Zaghouani, editors, *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 138–146, Valencia, Spain, April 2017. Association for Computational Linguistics. doi:10.18653/v1/W17-1317.

M. Bousri, R. Bensalem, S. Bessa, Z. Lamri, C. Zakaria, and N. Bousbia. Rumor detection in algerian arabizi based on deep learning and associations. In *Modeling and Implementation of Complex Systems*, Lecture Notes in Networks and Systems, pages 165–176. Springer, Cham, October 2022. doi:10.1007/978-3-031-18516-8_12. URL `https://doi.org/10.1007/978-3-031-18516-8_12`. In: Complex Systems Series.

T Britannica. Arabic language. In *t*. https://www.britannica.com/topic/Arabic-language, USA, 11 2023.

Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31:102–107, 2016. doi:10.1109/MIS.2016.31.

José Gabriel Carrasco Ramírez. Natural language processing advancements: Breaking barriers in human-computer interaction. *Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023*, 3(1):31–39, Feb. 2024. doi:10.60087/jaigs.v3i1.63. URL `https://ojs.boulibrary.com/index.php/JAIGS/article/view/63`.

Amina Chader, Lina Dihia, Lamine Hamdad, Merieme Belkheir, and Wissam Hennoune. Sentiment analysis for arabizi: Application to algerian dialect. In *Proceedings*, pages 475–482, 2019. doi:10.5220/0008353904750482.

Jingfeng Cui, Zhaoxia Wang, Seng Beng Ho, and Erik Cambria. Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, 56 (8):8469–8510, August 2023. doi:10.1007/s10462-022-10386-z.

Abdelghani Dahou, Abdelhalim Hafedh Dahou, Mohamed Amine Cheragui, Amin Abdedaiem, Mohammed A. A. Al-qaness, Mohamed Abd Elaziz, Ahmed A. Ewees, and Zhonglong Zheng. A survey on dialect arabic processing and analysis: Recent advances and future trends. *ACM Transactions on Asian and Low-Resource Language Information Processing*, July 2025. ISSN 2375-4699. doi:10.1145/3747290.

Nhan Cach Dang, María N. Moreno-García, and Fernando De la Prieta. Sentiment analysis based on deep learning: A comparative study. *Electronics (Switzerland)*, 9, 3 2020. ISSN 20799292. doi:10.3390/electronics9030483.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

Monireh Ebrahimi, Amir Hossein Yazdavar, and Amit Sheth. Challenges of sentiment analysis for dynamic events. *IEEE Intelligent Systems*, 32(5):70–75, September 2017. doi:10.1109/MIS.2017.3711649.

Ibrahim Abu El-Khair. Effects of stop words elimination for arabic information retrieval: A comparative study, 2017. URL `https://arxiv.org/abs/1702.01925`.

Nasrin Elhassan, Giuseppe Varone, Rami Ahmed, Mandar Gogate, Kia Dashtipour, Hani Almoamari, Mohammed A. El-Affendi, Bassam Naji Al-Tamimi, Faisal Albalwy, and Amir Hussain. Arabic sentiment analysis based on word embeddings and deep learning. *Computers*, 12(6):126, 2023a. ISSN 2073-431X. doi:10.3390/computers12060126.

Nasrin Elhassan, Giuseppe Varone, Rami Ahmed, Mandar Gogate, Kia Dashtipour, Hani Almoamari, Mohammed A. El-Affendi, Bassam Naji Al-Tamimi, Faisal Albalwy, and Amir Hussain. Arabic sentiment analysis based on word embeddings and deep learning. *Computers*, 12(6), 2023b. ISSN 2073-431X. doi:10.3390/computers12060126.

Mostafa Elmasry, Taysir Soliman, and Abdel-Rahman Hedar. Sentiment analysis of arabic slang comments on facebook. *Int. J. Comput. Technol.*, 12(5):3470–3478, January 2014. doi:10.24297/ijct.v12i5.2917.

Ashraf Elnagar, Sane Yagi, Youssef Mansour, Leena Lulu, and Shehdeh Fareh. A benchmark for evaluating arabic contextualized word embedding models. *Information Processing & Management*, 60(5):103452, 2023. ISSN 0306-4573. doi:10.1016/j.ipm.2023.103452.

Ibrahim Abu Farha and Walid Magdy. Mazajak: An online arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-4621.

Ibrahim Abu Farha and Walid Magdy. A comparative study of effective approaches for arabic sentiment analysis. *Information Processing and Management*, 58(2):102438, March 2021. doi:10.1016/j.ipm.2020.102438.

Aref Farhadipour, Hossein Ranjbar, Masoumeh Chapariniya, Teodora Vukovic, Sarah Ebling, and Volker Dellwo. Multimodal emotion recognition and sentiment analysis in multi-party conversation contexts. *arXiv preprint arXiv:2503.06805*, 2025. doi:10.48550/arXiv.2503.06805.

Ronen Feldman. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, apr 2013. ISSN 0001-0782. URL `https://doi.org/10.1145/2436256.2436274`.

Raheela Firdaus, Fakhra Yasmin, Syed Arslan Haider, Sarwat Jahan, and Petra Poulova. Emotion detection using machine learning: A study on human computer interaction. *Journal of Artificial Intelligence and Technology*, 5:125–134, Mar. 2025. doi:10.37965/jait.2025.0708.

Chayma Fourati, Hatem Haddad, Abir Messaoudi, Moez BenHajhmida, Aymen Ben Elhaj Mabrouk, and Malek Naski. Introducing a large tunisian arabizi dialectal dataset for sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 226–230, Kyiv, Ukraine (Virtual), 2021. Association for Computational Linguistics. `https://aclanthology.org/2021.wanlp-1.25`.

Yao Fu, Qiong Liu, Qing Song, Pengzhou Zhang, and Gongdong Liao. Multi-hm: A chinese multimodal dataset and fusion framework for emotion recognition in human–machine dialogue systems. *Applied Sciences*, 15(8), 2025. ISSN 2076-3417. doi:10.3390/app15084509. URL `https://www.mdpi.com/2076-3417/15/8/4509`.

K. Gaanoun, A. M. Naira, A. Allak, and I. Benelallam. Darijabert: a step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, 20:2463–2468, January 2024. doi:10.1007/s41060-023-00498-2.

Souvick Ghosh. Chapter 9 - sentiment-aware design of human–computer interactions: How research in human–computer interaction and sentiment analysis can lead to more user-centered systems? In Dipankar Das, Anup Kumar Kolya, Abhishek Basu, and Soham Sarkar, editors, *Computational Intelligence Applications for Text and Sentiment Data Analysis*, Hybrid Computational Intelligence for Pattern Analysis and Understanding, pages 209–224. Academic Press, 2023. ISBN 978-0-323-90535-0. doi:https://doi.org/10.1016/B978-0-32-390535-0.00014-8. URL `https://www.sciencedirect.com/science/article/pii/B9780323905350000148`.

S. Hajbi, O. Amezian, N. El Moukhi, R. Korchiyne, and C. Younes. Moroccan arabizi-to-arabic conversion using rule-based transliteration and weighted levenshtein algorithm. *Scientific African*, 23, March 2024. doi:10.1016/j.sciaf.2024.e02073.

Makrufa Hajirahimova and Marziya Ismayilova. Machine learning-based sentiment analysis of twitter data. *Problems of Information Society*, 13(1):52–60, January 2022. ISSN 2077-964X. doi:10.25045/jpis.v13.i1.07.

Shatha Ali A. Hakami, Robert Hendley, and Phillip Smith. Arabic emoji sentiment lexicon (Arab-ESL): A comparison between Arabic and European emoji sentiment lexicons. In Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors, *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 60–71, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics. `https://aclanthology.org/2021.wanlp-1.7/`.

Salima Harrat, Karima Meftouh, Mourad Abbas, Khaled-Walid Hidouci, and Kamel Smaili. An algerian dialect: Study and resources. *International Journal of Advanced Computer Science and Applications*, 7(3), 2016. doi:10.14569/IJACSA.2016.070353.

Hassan Hazimeh, Ali Takach, Hadi Harb, Karim Daher, Ezzat Mugellini, and Omar Abou Khaled. Saad: Sentiment analysis and abuse detection for arabizi language on social media. *Journal of Electrical Systems*, 20:2463–2468, September 2024.

Karishma Hegde and Hemadri Jayalath. Emotions in the loop: A survey of affective computing for emotional support, 2025. URL `https://arxiv.org/abs/2505.01542`.

Fatemah Husain, Hana Al-Ostad, and Halima Omar. A weak supervised transfer learning approach for sentiment analysis to the kuwaiti dialect. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 161–173, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.wanlp-1.15.

Sabrina Iarlori, Andrea Monteriú, David Perpetuini, Chiara Filippini, and Daniela Cardone. Editorial: Affective computing and mental workload assessment to enhance human-machine interaction. *Frontiers in Neuroergonomics*, 5:1412744, 2024. doi:10.3389/fnrgo.2024.1412744.

Samawel Jaballi, Salah Zrigui, Mohamed Sghaier, Dhaou Berchech, and Mounir Zrigui. Sentiment analysis of tunisian users on social networks: Overcoming the challenge of

multilingual comments in the tunisian dialect. pages 176–192, September 2022. ISBN 978-3-031-16013-4. doi:10.1007/978-3-031-16014-1_15.

Mouad Jbel, Mourad Jabrane, Imad Hafidi, and Abdulmutallib Metrane. Sentiment analysis dataset in moroccan dialect: bridging the gap between arabic and latin scripted dialect. *Language Resources and Evaluation*, September 2024. ISSN 1574-0218. doi:10.1007/s10579-024-09764-6.

Abdelrahman Kaseb and Mona Farouk. Saids: A novel approach for sentiment analysis informed of dialect and sarcasm, 2023. URL `https://arxiv.org/abs/2301.02521`.

Amit Khan, Dipankar Majumdar, and Bikromadittya Mondal. Sentiment analysis of emoji fused reviews using machine learning and bert. *Scientific Reports*, 15(1):7538, 2025. doi:10.1038/s41598-025-92286-0.

Barbara Kitchenham and Pearl Brereton. A systematic review of systematic review process research in software engineering. *Inf. Softw. Technol.*, 55(12):2049–2075, 2013. ISSN 0950-5849. doi:10.1016/j.infsof.2013.07.010.

Badia Klouche, Sidi Mohamed Benslimane, and Rim Bennabi. Ooredoo rayek: A business decision support system based on multi-language sentiment analysis of algerian operator telephones. *Int. J. Technol. Diffus.*, 11(2):66–81, February 2020. ISSN 1947-9301. doi:10.4018/IJTD.2020040105.

Mahander Kumar, Lal Khan, and Hsien-Tsung Chang. Evolving techniques in sentiment analysis: A comprehensive review. *PeerJ Comput. Sci.*, 11:e2592, January 2025. doi:10.7717/peerj-cs.2592.

Yuki Kuwabara and Yu Suzuki. Automatic stopword generation based on attention for document classification using neural networks. *Journal of Information Processing*, 32: 487–495, 05 2024. doi:10.2197/ipsjjip.32.487.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. Lstm-cnn deep learning model for sentiment analysis of dialectal arabic. In Kamel Smaïli, editor, *Arabic Language Processing: From Theory to Practice*, pages 108–121, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32959-4.

Raki Lachraf, El Moatez Billah Nagoudi, Youcef Ayachi, Ahmed Abdelali, and Didier Schwab. ArbEngVec : Arabic-English cross-lingual word embedding model. In Wassim El-Hajj, Lamia Hadrich Belguith, Fethi Bougares, Walid Magdy, Imed Zitouni, Nadi Tomeh, Mahmoud El-Haj, and Wajdi Zaghouani, editors, *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 40–48, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-4605.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977. ISSN 0006-341X. doi:10.2307/2529310.

Ruijun Liu, Yuqian Shi, Changjiang Ji, and Ming Jia. A survey of sentiment analysis based on transfer learning. *IEEE access*, 7:85401–85412, 2019.

Anudeep V. Mandalam, Omar Oueslati, Erik Cambria, and Yogesh Sharma. Arabizivec: A set of arabizi word embeddings for informal arabic sentiment analysis, 2024. Preprint.

Abir Masmoudi, Jamila Hamdi, and Lamia Hadrich Belguith. Deep learning for sentiment analysis of tunisian dialect. *Computación y Sistemas*, 25(1), February 2021. doi:10.13053/cys-25-1-3472.

M'hamed Mataoui, Omar Zelmati, and Madiha Boumechache. A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Res. Comput. Sci.*, 110 (1):55–70, December 2016. ISSN 1870-4069. doi:10.13053/rcs-110-1-5.

Yassir Matrane, Faouzia Benabbou, and Nawal Sael. A systematic literature review of arabic dialect sentiment analysis. *Journal of King Saud University - Computer and Information Sciences*, 35(6):101570, 2023a. ISSN 1319-1578. doi:10.1016/j.jksuci.2023.101570.

Yassir Matrane, Faouzia Benabbou, and Nawal Sael. A systematic literature review of arabic dialect sentiment analysis. *Journal of King Saud University - Computer and Information Sciences*, 35(6):101570, 2023b. ISSN 1319-1578. doi:10.1016/j.jksuci.2023.101570.

Yassir Matrane, Faouzia Benabbou, and Zineb Ellaky. Enhancing moroccan dialect sentiment analysis through optimized preprocessing and transfer learning techniques. *IEEE Access*, 12:187756–187777, 2024. doi:10.1109/ACCESS.2024.3514934.

Yassir Matrane, Faouzia Benabbou, Zineb Ellaky, and Chaimae Zaoui. *An Automatic Stop Words Removal in Maghrebi Arabic Dialect Text Classification Using Part of Speech Tagging*, page pp 187–196. springer, 05 2025. ISBN 978-3-031-88653-9. doi:10.1007/978-3-031-88653-9_19.

Annavarapu Chandra Sekhara Rao Mayur Wankhade and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55:5731–5780, 10 2022. ISSN 15737462. doi:10.1007/s10462-022-10144-1.

Ahmed Cherif Mazari and Abdelhamid Djeffal. Deep learning-based sentiment analysis of algerian dialect during hirak 2019. In *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-Being (IHSH 2020)*, pages 233–236, Boumerdes, Algeria, February 2021. Institute of Electrical and Electronics Engineers Inc. ISBN 9781665440844. doi:10.1109/IHSH51661.2021.9378753.

Ahmed Cherif Mazari and Abdelhamid Djeffal. Sentiment analysis of algerian dialect using machine learning and deep learning with word2vec. *Informatica (Slovenia)*, 46 (6):67–78, 2022a. ISSN 1854-3871. doi:10.31449/inf.v46i6.3340.

Ahmed Cherif Mazari and Abdelhamid Djeffal. Sentiment analysis of algerian dialect using machine learning and deep learning with word2vec. *Informatica*, 46(6):81–96, 2022b. doi:10.31449/inf.v46i6.3340.

Walaa Medhat, Ahmed H. Yousef, and Hoda Korashy. Egyptian dialect stopword list generation from social network data, 2015. URL `https://arxiv.org/abs/1508.02060`.

Freha Mezzoudj, Mourad Loukam, and Fatma Zohra Belkredim. Arabic algerian oranee dialectal language modelling oriented topic. *International Journal of Informatics and Applied Mathematics*, 2(2):1–14, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. https://arxiv.org/abs/1301.3781.

Ammar Mohammed and Rania Kora. Deep learning approaches for arabic sentiment analysis. *Social Network Analysis and Mining*, 9(1):52, December 2019. ISSN 1869-5450. doi:10.1007/s13278-019-0596-4.

G. Mostafa, Md Sabbir Ahamed, and Md Kishor Morol. Improve the sentiment of bengali language texts with stopword removal. In *Proceedings of the 3rd International Conference on Computing Advancements*, ICCA '24, pages 414–422, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713828. doi:10.1145/3723178.3723233.

Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. An algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1202–1210, Marseille, France, May 2020. European Language Resources Association. ISBN 9791095546344. `https://aclanthology.org/2020.lrec-1.151`.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. Afrisenti: A twitter sentiment analysis benchmark for african languages, February 2023.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1299.

D. Namly, K. Bouzoubaa, and A. Yousfi. A bi-technical analysis for arabic stop-words detection. *COMPUSOFT: An International Journal of Advanced Computer Technology*, 8(05):3126–3134, 2024. URL `https://ijact.in/index.php/j/article/view/491`.

Zineb Nassr, Nawal Sael, and Faouzia Benabbou. Preprocessing arabic dialect for sentiment mining: State of art. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLIV-4/W3 of *ISPRS Archives*, pages 323–330. Copernicus Publications, 2020. doi:10.5194/isprs-archives-XLIV-4-W3-2020-323-2020.

Zineb NASSR, Nawal SAEL, and Faouzia BENABBOU. Generate a list of stop words in moroccan dialect from social network data using word embedding. In *2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)*, pages 66–73, 2021. doi:10.1109/ICDATA52997.2021.00022.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille,

France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.868.

Mansour Omar, Aboelela Eman, Talaat Remon, and Bustami Mahmoud. Transformer-based ensemble model for dialectal arabic sentiment classification. *PeerJ Computer Science*, 11:e2644, 2025. doi:10.7717/peerj-cs.2644.

Abubakr H. Ombabi, Wael Ouarda, and Adel M. Alimi. Deep learning cnn–lstm framework for arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, 10(1):53, December 2020. doi:10.1007/s13278-020-00668-1.

Sara Ouahabi, El Ouahabi Safâa, and El Wardani Dadi. Multi-domain dataset for moroccan arabic dialect sentiment analysis in social networks. pages 01–07, 11 2023. doi:10.1109/CloudTech58737.2023.10366119.

L. Ouchene and S. Bessou. Sentiment analysis for algerian dialect tweets. In *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 235–239, Al-Khobar, Saudi Arabia, 2022. doi:10.1109/CICN56167.2022.10008314.

L. Ouchene and S. Bessou. Fasttext embedding and lstm for sentiment analysis: An empirical study on algerian tweets. In *2023 International Conference on Information Technology (ICIT)*, pages 51–55, Amman, Jordan, 2023. doi:10.1109/ICIT58056.2023.10226060.

Oumaima Oueslati, Erik Cambria, Moez Ben HajHmida, and Habib Ounelli. A review of sentiment analysis research in arabic language. *Future Gener. Comput. Syst.*, 112: 408–430, November 2020. doi:10.1016/j.future.2020.05.034.

Yuxing Qi and Zahratu Shabrina. Sentiment analysis using twitter data: a comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, 13(1):31, 2023. doi:10.1007/s13278-023-01030-x.

Hichem Rahab, Abdelhafid Zitouni, and Mahieddine Djoudi. Sana: Sentiment analysis on newspapers comments in algeria. *J. King Saud Univ. Comput. Inf. Sci.*, 33(7):899–907, September 2021. ISSN 2213-1248. doi:10.1016/j.jksuci.2019.04.012.

Razzia Rahmoun-Mrabet. A sociolinguistic study of the outcomes of arabic-french contact in the algerian dialect tlemcen speech community as a case study. *International Journal of Cognitive and Language Sciences*, 12(11):1547–1554, 2018. ISSN 1307-6892. https://publications.waset.org/vol/143.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics. doi:10.18653/v1/S17-2088.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 810–817,

Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). https://aclanthology.org/L14-1265/.

Dhai Eddine Salhi, Abdelkamel Tari, and Tahar Kechadi. Sentiment analysis application on twitter for e-reputation. In *Proceedings of the Sixth International Conference on the Image and Signal Processing and their Applications*, Mostaganem, Algeria, November 2019. IEEE. doi:10.1109/ISPA48434.2019.8966833.

Tengku Mohd Tengku Sembok and Belal Mustafa Abuata. Arabic stop words for information retrieval systems. *International Journal of Religion*, 6(1):121–127, 2025. doi:10.61707/k7rnm813.

Arwa A. Al Shamsi and Sherief Abdallah. Text mining techniques for sentiment analysis of arabic dialects: Literature review. *Advances in Science, Technology and Engineering Systems Journal*, 6:1012–1023, 2 2021. ISSN 24156698. doi:10.25046/aj0601112.

Ahmed Sherif and Caroline Sabty. Sentiment analysis for egyptian arabic-english code-switched data using traditional neural models and advanced language models. In *Speech and Computer: 26th International Conference, SPECOM 2024, Belgrade, Serbia, November 25–28, 2024, Proceedings, Part II*, page 54–69, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-78013-4. doi:10.1007/978-3-031-78014-1_5.

Alexander Shevtsov, Maria Oikonomidou, Despoina Antonakaki, Polyvios Pratikakis, and Sotiris Ioannidis. What tweets and youtube comments have in common? sentiment and graph analysis on data related to us elections 2020. *PLOS ONE*, 18(1):1–31, 01 2023. doi:10.1371/journal.pone.0270542.

Zhiqiang Shi and Ruchit Agrawal. A comprehensive survey of contemporary arabic sentiment analysis: Methods, challenges, and future directions, 2025.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 9 2017. ISSN 02628856. doi:10.1016/j.imavis.2017.08.003.

Abu Bakr Soliman, Kareem Eisa, and Samhaa R. El-Beltagy. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265, 2017a. doi:10.1016/j.procs.2017.10.117.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265, 2017b. ISSN 1877-0509. doi:10.1016/j.procs.2017.10.117.

Jindi Song. A review of the application of natural language processing in human-computer interaction. *Applied and Computational Engineering*, 106:111–117, 11 2024. doi:10.54254/2755-2721/106/20241328.

Assia Soumeur, Mheni Mokdadi, Ahmed Guessoum, and Amina Daoud. Sentiment analysis of users on social networks: Overcoming the challenge of the loose usages of the algerian dialect. *Procedia Comput. Sci.*, 142:26–37, 2018. ISSN 1877-0509. doi:10.1016/j.procs.2018.10.458.

Jing Su, Qun Chen, Yanyan Wang, Lijun Zhang, Wei Pan, and Zhanhuai Li. Sentence-level sentiment analysis based on supervised gradual machine learning. *Scientific Reports*, 13(1):14500, 2023. doi:10.1038/s41598-023-41485-8.

T. Tobaili, M. Fernandez, H. Alani, S. Sharafeddine, H. Hajj, and G. Glavaš. Senzi: A sentiment analysis lexicon for the latinised arabic (arabizi). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 1203–1211, Varna, Bulgaria, September 2019. Association for Computational Linguistics. doi:10.26615/978-954-452-056-4_138.

Said Touileb. Nerdz: A preliminary dataset of named entities for algerian. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) and the 12th International Joint Conference on Natural Language Processing (IJCNLP), Volume 2: Short Papers*, pages 95–101, Online, 2022. doi:10.18653/v1/2022.aacl-short.13.

Samia Touileb and Jeremy Barnes. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3700–3712, Online, August 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.findings-acl.324.

Ciprian-Octavian Truică, Elena-Simona Apostol, Maria-Luiza Erban, and Adrian Paschke. Topic-based document-level sentiment analysis using contextual cues. *Mathematics*, 9:2722, 10 2021. ISSN 2227-7390. doi:10.3390/math9212722.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL `https://arxiv.org/abs/1706.03762`.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-trained language models and their applications. *Engineering*, 25:51–65, 6 2023. ISSN 20958099. doi:10.1016/j.eng.2022.04.024.

Yili Wang, Jiaxuan Guo, Chengsheng Yuan, and Baozhu Li. Sentiment analysis of twitter data, 11 2022. ISSN 20763417.

Asma Al Wazrah and Sarah Alhumoud. Sentiment analysis using stacked gated recurrent unit for arabic tweets. *IEEE Access*, 9:137176–137187, 2021. doi:10.1109/ACCESS.2021.3114313.

Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53:4335–4385, 8 2020. ISSN 15737462. doi:10.1007/s10462-019-09794-5.

Taha Zerrouki. Tashaphyne: A python package for arabic light stemming. *Journal of Open Source Software*, 9(93):6063, 2024. doi:10.21105/joss.06063.

Tianyu Zhang and Fucheng You. Research on short text classification based on textcnn. *Journal of Physics: Conference Series*, 1757:012092, 01 2021. doi:10.1088/1742-6596/1757/1/012092.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges, 3 2022. URL `http://arxiv.org/abs/2203.01054`.

David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation, 4 2018. ISSN 21586578.

Appendix A

Annotation Guidelines

Tweets were annotated according to the overall sentiment intended by the author, rather than the literal meanings of individual words. This approach was crucial for correctly identifying figurative language, sarcasm, and emotionally ambiguous expressions common in Algerian Arabic social media. The following special cases were considered during annotation:

- Annotators used tone and context to differentiate literal from sarcastic expressions.
- Supplications and prayers were generally labeled as neutral, unless the context indicated a strong positive or negative sentiment.
- Tweets containing a mix of Algerian dialect, MSA, French, or Arabizi were annotated based on the overall sentiment.

A.0.1 Annotator Disagreements

Some tweets were challenging to interpret, leading to initial disagreements often caused by sarcasm, negation, or contradictory tones. Table A.1 shows examples of such cases. Disagreements were resolved through discussion or majority voting.

A.0.2 Inter-Annotator Agreement

To assess annotation consistency, a subset of 1,000 tweets was independently annotated by all three annotators. Cohen's Kappa (Landis and Koch, 1977) was computed for each annotator pair and averaged. The overall score was 0.83, indicating substantial agreement (Landis and Koch, 1977).

Table A.1 Sample Tweets illustrating Annotation Disagreements for Algerian Sentiment Analysis

Tweet	A1	A2	A3	Final Label	Description
في الجزائر كل شيئ ممكن In Algeria, anything is possible	Pos	Neg	Neg	Neg	Irony
تافه بصح هايلhhhh Silly but great, hhhh	Pos	Neg	Neg	Neg	Sarcasm
هذه انترنت الجزائر This is Algerian internet	Neu	Neg	Neu	Neg	Irony
😂 Legends والله Legends, I swear	Pos	Neg	Neg	Neg	Sarcasm
مرحبا بك في تويتر و في الجزائر الجديده😂 Welcome to Twitter and to the new Algeria	Pos	Neg	Neg	Neg	Irony
في راسك ان شاء الله، الجزائر فيها رجال May God keep you safe, Algeria has real men	Neg	Neg	Pos	Pos	Mixed tone
🤔 المنشط ياربي قطعتلي قلبي pravooo The host, you broke my heart, bravo	Neg	Neg	Pos	Pos	Mixed tone

Appendix B

Arabizi to Arabic Mapping

Table B.1 presents selected examples from our Arabizi normalization mapping file used during preprocessing.

Table B.1 Sample Arabizi-to-Arabic mapping entries used for Normalization

Arabic Word	Arabizi Variants
والله I swear to God	wallahi, woulah, walah, wallah, oulah, wllh, wlh
شكرا Thank you	choukran, chokran, chokrane
كثير Many	bezzaf, bezaf, bzef, bezef, bzzaffff
يعمل He does	ydirou, ydiro, ndirou
اخي My brother	5oya, khoya
حبيبي My dear	7abibi, habibi, hbibi
ربما Maybe	wakila, wakil
كيف How	kifach, kifech
لابأس It's okay	m3lich, me3lich, mehlich, maalich, merlich, ma3lich
اذهب Go	rouhi, rouh, rouhou, nrouh, roho, roh
ان شاءالله God willing	inchallahe, nchallah, anchallah, inchaallah, inchalah, inchala, inchala, challah, inchaellah, inchaellah, inchaalah, inchaa, incheallah, inshallah,inshaallah
الحمدلله Praise be to God	hamdoullillah, hamdoulilah, alhamdoulilah, hamdoulah, elhamdoulilah, hamdullilah, hamdoullah, hamdoulillah, hamdoulill, hamdoulileh,hamdouleh, hamdulilah, alhamdoulillah, alhamdulilah, hamdulillah, hamdullah