**People's Democratic Republic of Algeria**
**Ministry of Higher Education and Scientific Research**

**Kasdi Merbah University of Ouargla**
**Faculty of New Technologies of Information and Communication**
**Department of Computer Science and Information Technology**


# *Doctoral Thesis*

Thesis submitted in partial fulfillment of the requirements for the degree of PhD 3rd cycle in computer science.

**Option** :

Artificial Intelligence

---

# A Transfer Learning-Based Approach for Arabic Sentiment Analysis

---

**Presented by**

*M^r Mohammed Elsadiq BARMATI*

**Jury members**

| | | |
|---|---|---|
| **President:** | *Ahmed KORICHI* | Professor, University of Ouargla, Algeria |
| **Supervisor:** | *Bachir SAID* | M.C.A, University of Ouargla, Algeria |
| **Examiner:** | *Oussama AIADI* | Professor, University of Ouargla, Algeria |
| **Examiner:** | *Hocine MERABTI* | M.C.A, University of Ouargla, Algeria |
| **Examiner:** | *Mounir BEGGAS* | M.C.A, University of El-Oued, Algeria |
| **Examiner:** | *Said GADRI* | M.C.A, University of M'sila, Algeria |

**2025/2026**

For my parents.
For my little family.
For my sisters and brothers.

# Acknowledgements

First and foremost, all praise and gratitude are due to Allah, the Almighty, for granting me the strength, perseverance, and opportunity to complete this research. His guidance and blessings have been instrumental throughout my academic journey, and without his blessings, this achievement would not have been possible.

I would like to express my deepest gratitude to my supervisor, BACHIR Said, for his unwavering guidance, invaluable insights, and continuous support throughout the course of this research. His mentorship has been instrumental in shaping both the academic and personal dimensions of this dissertation.

Heartfelt thanks are also directed to the distinguished jury members — Prof. Ahmed KORICHI, Dr. Oussama AIADI, Dr. Hocine MERABTI, Dr. Mounir BEGGAS, and Dr. Said GADRI for generously dedicating their time and expertise to the evaluation of this work and for their constructive observations that enhanced its academic merit.

With deep affection and enduring respect, I dedicate this work to my father, Abdelkarim, whose strength, guidance, and unwavering encouragement have been a constant source of motivation. Above all that, he has been a truly good and caring father, for which I am deeply grateful.

I also honour the memory of my late mother, whose love, sacrifice, and lasting influence continue to inspire all that I strive to accomplish. My heartfelt gratitude extends as well to my stepmother, whose kindness, support, and presence have been a meaningful part of this journey.

To my beloved wife and our little family, I owe my deepest love and thanks. Your patience, support, and faith in me have sustained me through every challenge and given true meaning to this journey. Your presence has been my greatest comfort and motivation.

I extend heartfelt thanks to my brothers and sisters for their support, encouragement, and belief in me. Their companionship and solidarity have been a pillar of strength throughout this endeavor.

Lastly, I am sincerely thankful to all those who have supported me in any form, my friends, my colleagues, and my mentors. Each contribution, no matter how small, has played a part in making this work possible.

# ملخص

يُعد تحليل المشاعر في اللغة العربية من المهام الجوهرية في معالجة اللغة الطبيعية، إذ يتطلب تحديد وتصنيف المشاعر في النصوص العربية. وتكمن تعقيداته في السمات اللغوية الغنية للعربية، مثل الاشتقاق الصرفي المعقد، وتنوع اللهجات، والاستخدام الشائع للتعبيرات المجازية كالسخرية. هذه التحديات تحدّ من قدرة النماذج التقليدية على الأداء الدقيق.

في هذا السياق، تقدم هذه الرسالة إطارين مبتكرين يعتمدان على التعلم بالنقل لتحسين أداء تحليل المشاعر.

الإطار الأول متعدد الوسائط، حيث يدمج بين التمثيلات النصية المشتقة من نماذج التحويل العربية المدربة مسبقًا، وميزات عددية وتصنيفية غير متجانسة، ما يعزز التنبؤ بالمشاعر عبر دمج مصادر متعددة للمعلومات، كما أثبتت نتائجه على بيانات اللهجة الشامية.

أما الإطار الثاني، فيستخدم التعلم متعدد المهام لتنفيذ ثلاث مهام مترابطة: تصنيف المشاعر، واكتشاف السخرية، وتحديد اللهجة، باستخدام مشفر مشترك قائم على المحولات وفك ترميز خاص بكل مهمة. أظهرت التجارب تفوق هذا النموذج على النماذج أحادية المهمة، بفضل استغلاله للمعرفة المشتركة بين المهام المذكوره.

تُسهم هاتان المعماريتان في بناء حل موحد وقابل للتوسّع لتحليل المشاعر بالعربية، وتختتم الرسالة بمقترحات بحثية مستقبلية تشمل نمذجة المهام هرميًا، وتطوير آليات انتباه مخصصة، ودمج الميزات الجدولية ضمن تصميمات التعلم متعدد المهام.

**الكلمات المفتاحية:** تحليل المشاعر باللغة العربية، اكتشاف السخرية في اللغة العربية، التعلم متعدد المهام، البيانات متعددة الأنماط، التعلم الآلي، التعلم بالنقل، التعلم العميق.

# Abstract

Arabic sentiment analysis is a critical task in natural language processing (NLP) that involves identifying and classifying sentiments expressed in Arabic text. Its complexity arises from the language's rich morphology, widespread dialectal variation, and frequent use of figurative expressions such as sarcasm. Although traditional machine learning models have contributed to the field of NLP, they often fail to capture these linguistic subtleties, limiting their effectiveness in real-world applications.

This dissertation introduces two novel transfer learning architectures to advance Arabic sentiment analysis: a multimodal framework and a multi-task learning (MTL) framework. The first architecture combines textual representations from pre-trained Arabic transformer models with tabular categorical and numerical features to form a multimodal input. Evaluated on the ArSenTD-Lev dataset, this approach demonstrates that incorporating heterogeneous modalities enhances sentiment classification performance.

The second architecture employs an MTL strategy that simultaneously performs sentiment classification, sarcasm detection, and dialect identification using a shared transformer encoder and task-specific decoders. By leveraging shared contextual knowledge and inter-task Interrelatedness, the MTL model enhances generalization and reduces overfitting. Empirical results across benchmark datasets, including ArSarcasm, ArSenTD-Lev, ASTD, and NADI, validate the superior performance of the proposed MTL model over conventional single-task baselines.

Collectively, the multimodal and MTL frameworks contribute to a unified and scalable solution for Arabic sentiment analysis. The dissertation concludes by outlining future research directions, including hierarchical task modeling, task-specific attention mechanisms, and the direct integration of tabular features into MTL architectures to further enhance task interaction and model interpretability.

**Keywords:** Arabic sentiment analysis, Arabic sarcasm detection, Multi-task learning, Multimodal data, machine learning, transfer learning, Deep learning.

# Résumé

L'analyse des sentiments en arabe constitue une tâche essentielle du traitement automatique du langage naturel (TALN), consistant à identifier et à classifier les émotions exprimées dans les textes arabes. Sa complexité découle de la richesse morphologique de la langue, de la grande diversité dialectale et de l'usage fréquent d'expressions figuratives telles que le sarcasme. Les modèles traditionnels d'apprentissage automatique peinent souvent à saisir ces subtilités linguistiques, limitant ainsi leur efficacité dans des contextes d'application réels.

Cette thèse propose deux architectures novatrices basées sur l'apprentissage par transfert pour faire progresser l'analyse des sentiments en arabe : un cadre multimodal et un cadre d'apprentissage multi-tâches (MTL).

La première architecture intègre des représentations textuelles extraites de modèles transformeurs pré-entraînés en arabe avec des caractéristiques tabulaires catégorielles et numériques, constituant une entrée multimodale. Évaluée sur le jeu de données ArSenTD-Lev, cette approche démontre que l'inclusion de modalités hétérogènes améliore les performances de classification des sentiments.

La seconde architecture adopte une stratégie MTL permettant d'apprendre simultanément la classification des sentiments, la détection du sarcasme et l'identification des dialectes à travers un encodeur transformeur partagé et des décodeurs spécifiques à chaque tâche. En tirant parti de connaissances contextuelles partagées et des synergies inter-tâches, le modèle MTL favorise une meilleure généralisation et atténue le surapprentissage. Les résultats empiriques sur les jeux de données ArSarcasm, ArSenTD-Lev, ASTD et NADI valident la supériorité du modèle MTL proposé par rapport aux approches mono-tâche classiques.

Les approches multimodale et MTL forment une solution unifiée pour l'analyse des sentiments en arabe. En conclusion, la thèse suggère des recherches futures, telles que la modélisation hiérarchique des tâches et l'intégration de données tabulaires pour améliorer l'interprétation et l'interaction entre tâches.

**Mots-clés :** Analyse de sentiment en arabe, Détection du sarcasme arabe Apprentissage multi-tâches, Données multimodales, Apprentissage automatique, Apprentissage par transfert , Apprentissage profond.

# Contents

# List of Figures

# List of Tables

# List of Publication

## Published Journal Papers

BARMATI Mohammed Elsadiq, SAID Bachir  Dahou Abdelghani. Multi-task learning for multi-dialect Arabic sentiment classification and sarcasm detection. Lang Resources  Evaluation 59, 2589–2612 (2025). https://doi.org/10.1007/s10579-025-09823-6. **Impact factor (2024): 1.80**

## Published Conference Papers

BARMATI Mohammed Elsadiq, SAID Bachir. Enhancing Arabic Sentiment Analysis through Multimodal-Toolkit: A Study on Levantine Arabic Dataset. The First National Conference for Applied Sciences and Engineering (NCASE-24) (2024).

# List of Abbreviations

**ANLP** . . . . . Arabic Natural Language Processing

**ArSenL** . . . . Arabic Sentiment Lexicon

**ArSenTD-Lev** Arabic Sentiment Twitter Dataset for the Levantine dialect

**ASD** . . . . . Arabic sarcasm detection

**ASA** . . . . . Arabic sentiment classification

**ASL** . . . . . Arabic Sentiment Lexicon

**ASTD** . . . . . Arabic Sentiment Tweets Dataset

**AWN** . . . . . Arabic WordNet

**BOW** . . . . . Bag-of-Words

**CNN** . . . . . Convolutional Neural Networks

**DAPT** . . . . . Domain-adaptive pretraining

**DT** . . . . . Decision Trees

**DWA** . . . . . Dynamic Weight Averaging

**FCL** . . . . . Fully Connected Layer

**GRUs** . . . . . Gated Recurrent Units

**KNN** . . . . . k-nearest neighbors

**LSTM** . . . . . Long Short-Term Memory network

**MSA** . . . . . Modern Standard Arabic

**MTL** . . . . . multi-task learning

**NADI** . . . . . Nuanced Arabic Dialect Identification

**NB** . . . . . Naïve Bayes

**NER** . . . . . Named Entity Recognition

**NLP** . . . . . natural language processing

**POS** . . . . . Part Of Speech

**PLMs** . . . . . Pretrained Language Models

**RNN** . . . . . Recurrent Neural Network

**RC** . . . . . Ridge Classifier

**SA** . . . . . Sentiment Analysis

**SAMA** . . . . . Standard Arabic Morphological Analyzer

**SC** . . . . . Sentiment Classification

**SD** . . . . . Sarcasm Detection

**STL** . . . . . single-task learning

**SVM** . . . . . Support Vector Machines

**TAPT** . . . . . Task-Adaptive Pretraining

**TALN** . . . . . traitemenT automatique du langage naturel

**TF-IDF** . . . . Term Frequency-Inverse Document Frequency

**X** . . . . . Twitter

# Chapter 1

# General introduction

## 1.1 Introduction

Traditionally, surveys have served as the principal methodology for eliciting public opinion on various issues. This process typically involves administering a structured questionnaire to a representative sample of individuals, allowing for a controlled exploration of their attitudes [1]. However, the rapid proliferation of internet access and the emergence of social media platforms such as X (formerly Twitter) and Facebook have transformed how opinions are generated, disseminated, and analyzed. These platforms have not only revolutionized interpersonal communication but also introduced new mechanisms for aggregating and interpreting public sentiment in real time.

This shift has led to the emergence of Sentiment Analysis (SA), the subject explored in this thesis. The aim of SA is to identify and classify opinions expressed in a text. Unlike traditional methods that rely on closed-ended survey instruments, sentiment analysis leverages vast volumes of unsolicited, user-generated content to assess public opinion dynamically. The technique is particularly applicable to social media texts, where users express sentiments about diverse topics ranging from political discourse and social issues to commercial products and services. Among social media platforms, X (Twitter) has become especially prominent due to its concise message format and real-time nature, offering a rich source of data for mining public attitudes. Sentiment analysis seeks to detect the polarity of such expressions, classifying them into categories such as positive, negative, or neutral.

Over the past decade, sentiment analysis has emerged as a prominent area of inquiry within both academic and commercial contexts. It has also become an integral component of data-driven decision-making in various industries. For instance, platforms such as LexisNexis [1] utilize sentiment analysis to monitor consumer perceptions and brand engagement by analyzing content from news outlets. Similarly, tools like IBM SPSS[2] provide quantitative sentiment summaries derived from large datasets to help organizations better understand customer preferences and behavioral trends. In the media sector, prominent news platforms including

[1] http://www.lexisnexis.com/risk/data-analytics.aspx

[2] http://www-01.ibm.com/software/analytics/spss/

Politico[3] and The Washington Post [4] employ sentiment analytics to present public opinion data concerning political figures and policy issues. Moreover, sentiment analysis plays a vital role in the financial sector, where institutions such as Wall Street integrate it into algorithmic trading systems. Technologies like OpFine [5] exemplify the application of sentiment-driven analytics to assess real-time financial developments and market movements [2].

The early applications of sentiment analysis focused primarily on product reviews, particularly those on e-commerce platforms such as Amazon [6], where star ratings provided easily quantifiable labels for supervised learning tasks. This facilitated the development of annotated datasets and spurred interest in sentiment detection across more complex text types, such as blogs, news articles, and online forums. As Twitter gained popularity, its real-time stream of public commentary became an invaluable resource for researchers, enabling applications ranging from commercial opinion tracking to public safety monitoring, such as earthquake detection [3]. Despite its utility, most sentiment analysis studies have historically focused on homogeneous sources like customer reviews, and only recently have models begun to generalize to more heterogeneous and nuanced forms of social media content.

Furthermore, while sentiment analysis has demonstrated substantial success in product evaluation and consumer feedback, its application to socially and politically charged discourse presents additional challenges. These include the presence of sarcasm, implicit sentiment, and the use of idiomatic or dialectal expressions all of which are common in social media communication. Consequently, questions remain regarding the appropriateness of conventional sentiment analysis techniques, originally designed for structured product reviews, for interpreting the complex affective signals embedded in informal digital interactions.

In the context of the Arabic language, sentiment analysis research initially lagged behind due to linguistic complexity, lack of resources, and the rich dialectal variation across the Arab world. However, recent years have witnessed a significant growth in scholarly interest and research output. Early efforts by Ahmad et al. [4] and Almas and Ahmad [5] applied grammatical analysis to sentiment classification in Arabic financial news articles, marking some of the first contributions to the field. Since then, the body of literature on Arabic sentiment analysis has expanded steadily, addressing various linguistic and computational challenges unique to the Arabic language.

## 1.2 Research Motivation

Sentiment analysis has attracted considerable attention across both academic and industrial domains, emerging as a vital tool for extracting and interpreting affective information from text. Significant advancements have been made in the development of sentiment analysis models. Nevertheless, the field remains an active and evolving area of research due to the linguistic diversity and complexity found across the

[3] `http://news.cnet.com/8301-13772`
[4] `http://www.washingtonpost.com/politics/mention-machine`
[5] `http://www.opfine.com/`
[6] `https://www.amazon.com/`

world's languages. Among these, in recent years, the Arabic-speaking population has demonstrated growing engagement with social media platforms especially X (Twitter) where users express opinions on a wide array of social, political, and cultural topics, including events. These developments provide public institutions and researchers with a valuable opportunity to apply sentiment analysis techniques to Arabic social media content, enabling the assessment of public response to government policies and social developments.

Despite the growth of Arabic sentiment analysis as a research area, the field continues to face substantial linguistic and technical challenges. Arabic is a morphologically rich and syntactically complex language, exhibiting a high degree of ambiguity, numerous irregular forms, and diverse dialectal variations that lack standardized orthographic conventions. These characteristics complicate natural language processing (NLP) tasks and limit the generalizability of models trained on Modern Standard Arabic (MSA) when applied to dialectal content. Furthermore, the scarcity of publicly available annotated resources and NLP tools for Arabic especially those targeting regional dialects significantly hampers progress in this domain, in contrast to the extensive resources available for English.

The overall objective of this research is motivated by the goal of developing a robust sentiment analysis framework. The focus is particularly on social media discourse written in dialectal Arabic, which often deviates from formal syntactic norms. Addressing this challenge is essential for producing accurate and contextually aware sentiment classification systems.

## 1.3 Research Objectives

The primary objective of this research is to enhance the performance and applicability of sentiment analysis in Arabic social media contexts, particularly by addressing the linguistic and computational challenges associated with dialectal Arabic. This study proposes the use of a multi-task learning (MTL) framework grounded in transformer-based models. This approach leverages the interrelation between sentiment classification and two linguistically relevant auxiliary tasks: sarcasm detection and dialect identification. Both tasks are highly pertinent to Arabic, where the presence of sarcasm and dialectal diversity can significantly alter the polarity and semantic interpretation of text. By training a shared model across these tasks, the research aims to exploit inter-task dependencies and improve generalization across different language varieties.

In addition to textual representations, this study also incorporates structured tabular features including categorical and numerical features using particular combining methods. The integration of such features is intended to enhance model interpretability, robustness to informal language, and adaptability to underrepresented dialects

The key research objectives are as follows:

- **Investigate the limitations of current Arabic sentiment analysis methods**, particularly in handling dialectal variation, informal expressions,

and the scarcity of annotated resources for social media texts.

- **Implement a multi-task learning (MTL) framework** that jointly performs sentiment classification, sarcasm detection, and dialect identification, thereby enhancing contextual understanding and improving task-specific generalization.
- **Evaluate and fine-tune transformer-based pre-trained language models** under both single-task and multi-task configurations, and assess their performance across various Arabic dialects.
- **Empirically compare the proposed MTL models with single-task baselines**, using both intrinsic (e.g., F1-score, accuracy) and extrinsic (e.g., robustness to dialects and sarcasm) evaluation metrics.
- **Assess the impact of incorporating dialect identification and sarcasm detection as auxiliary tasks** in improving the accuracy and generalizability of sentiment classification models on social media texts.
- **Integrate structured tabular features**, such as categorical and numerical features, with deep contextual embeddings to enrich the representation of Arabic social media text.
- **Validate and quantify the performance of the proposed architectures**, demonstrating their strengths and potential for real-world applications

## 1.4 Contributions

This dissertation presents several key contributions to the field of Arabic sentiment analysis, with a particular focus on leveraging multi-task learning and multimodal integration to address the challenges posed by dialectal variation, sarcasm, and limited resources in Arabic natural language processing (NLP). The contributions span theoretical advancements, methodological innovations, and empirical validations, and are organized as follows:

- **Combination of Tabular Features with Text feature:** This research introduces a new integration of structured tabular features comprising both categorical and numerical attributes with text. The combined representation enriches the model's understanding of Arabic social media content and improves classification robustness.
- **Development of a Multi-Corpus MTL Model for Dialectal Arabic Sentiment Analysis:** A new multi-task learning model is proposed that leverages multiple Arabic corpora. By employing transformer-based Arabic language models, this approach enhances classification performance across diverse dialects and domain-specific datasets, contributing to cross-corpus generalizability.

- **Comprehensive Evaluation of Single-Task and Multi-Task Transformer Models:** A systematic evaluation of both single-task and multi-task learning frameworks is conducted using transformer-based language models. This includes empirical comparisons across tasks such as sentiment analysis, sarcasm detection, and dialect identification to determine their individual and joint contributions.
- **Empirical Evidence for Task Interrelated between Sentiment and Sarcasm Detection:** Through detailed experimentation, this dissertation demonstrates that incorporating sarcasm detection as an auxiliary task can significantly improve sentiment classification performance in Arabic. This finding supports the hypothesis that sentiment polarity and sarcasm presence are interrelated phenomena that benefit from joint modeling.
- **Integration of Arabic Dialect Identification to Resolve Lexical Ambiguity:** This study incorporates Arabic dialect identification as an auxiliary task within the multi-task learning (MTL) framework to mitigate the semantic ambiguity caused by dialectal variation.

## 1.5 Thesis structure

This dissertation is organized into six main chapters, each systematically addressing the components of the research, from theoretical underpinnings to experimental evaluation and final conclusions. Apart from the opening chapter, which provides the general introduction, the structure of the dissertation is as follows:

- Chapter 2 presents the foundational background necessary for understanding the research. It begins with an overview of sentiment analysis, including its approaches and associated challenges. It then shifts focus to the complexities of processing Arabic, examining its linguistic characteristics, dialectal variation, and specific challenges in NLP. Special attention is given to Arabic transformers and the principles of multi-task learning (MTL), which constitute core components of the proposed methodology.
- Chapter 3 surveys prior research in Arabic sentiment analysis. It is divided into subsections covering lexicon-based approaches, traditional machine learning techniques, and recent advances in deep learning. It also discusses developments in transfer learning and multi-task learning for NLP, focusing on their applications to Arabic language processing. The chapter identifies gaps in the literature that this research aims to address.
- Chapter 4 presents the core contributions of the dissertation. It includes detailed descriptions of the proposed models and methodologies, such as integrating structured tabular features, dialect identification, and sarcasm detection tasks within a multi-task learning framework.

- Chapter 5 reports the experimental setup and results. It provides comprehensive details on datasets, preprocessing techniques, evaluation metrics, baseline comparisons, and experimental configurations. The performance of the proposed models is analyzed across various tasks, including combining methods, single-task and multi-task scenarios. Each experimental series is discussed in detail to evaluate the efficacy of auxiliary tasks and feature integration.
- Finally, chapter 6 summarizes the main findings of the research, discusses their implications for Arabic NLP, and reflects on the limitations encountered. It also outlines directions for future work, such as expanding dialect coverage, improving model generalizability, and incorporating multimodal and real-world deployment aspects.

# Chapter 2

# Overview of Sentiment Analysis and Arabic Natural Language Processing

## 2.1 Introduction

As Web technologies expand, the volume of user-generated content on social media platforms and web forums increases, necessitating advanced techniques like sentiment analysis (SA) to extract opinions from large textual datasets. SA involves Natural Language Processing (NLP) techniques to analyze and classify sentiments, distinguishing positive and negative views toward products and services [6, 7]. This task is particularly challenging for Arabic, which has a rich morphology, dialectal variations, and complex syntax. To address these challenges, transfer learning has proven to be a robust and scalable solution in Arabic language processing. By leveraging pre-trained models on high-resource languages, transfer learning enables the application of this knowledge to low-resource languages like Arabic, improving sentiment classification performance [8].

Moreover, multi-task learning (MTL) approaches have also enhanced SA for Arabic by enabling models to simultaneously learn related tasks. This shared learning improves the generalization and accuracy of the models[9]. These advancements in transfer learning and MTL have significantly improved sentiment analysis for Arabic, enabling more accurate sentiment extraction from social media and other platforms, thus enhancing applications in marketing, brand reputation management, and political sentiment tracking.

## 2.2 Sentiment analysis

Sentiment analysis (SA), also referred to as opinion mining, constitutes a pivotal task within the broader field of natural language processing (NLP). It involves identifying and extracting subjective information from text, thereby facilitating the assessment of opinions, emotions, and attitudes embedded in language. SA plays a crucial role across diverse applications, including social media monitoring, customer feedback

analysis, product review aggregation, and political discourse evaluation. The social media platforms such as X (Twitter), Facebook, and YouTube have heightened the need for automated SA systems capable of capturing public sentiment efficiently. These systems are increasingly influential in shaping decision-making processes across academic research, industry practices, and individual consumer behaviors [10].

SA can be utilised in differing aspects of society and interests for example it can be used in business, public concerns, commercial or in politics. Within the commercial sector, sentiment analysis is systematically employed to monitor brand reputation, evaluate customer satisfaction, and guide data-driven digital marketing strategies. Prominent platforms such as Google Product Search and TripAdvisor integrate automated sentiment analysis systems to process and synthesize user-generated reviews. These analytical outputs have a measurable impact on strategic business decisions and consumer purchasing patterns. Moreover, sentiment analysis within the public sector has been utilized as an example to evaluate societal attitudes toward vaccination during the COVID-19 pandemic, thereby supporting health authorities in mitigating misinformation and refining public communication initiatives [11].

Furthermore, SA has been applied to gauge public opinion in political contexts. Studies have employed SA techniques to analyze Arabic online texts, facilitating the understanding of public sentiment toward political events. These applications underscore the versatility of SA as a tool for real-time opinion tracking and strategic decision-making across diverse domains[12].

### 2.2.1 Sentiment analysis approaches

As illustrated in 2.1 there are three main approaches to the sentiment analysis process: Lexicon-Based, Machine Learning, and Hybrid approaches. Table 2.1 illustrates the strengths and weaknesses of sentiment analysis approaches.

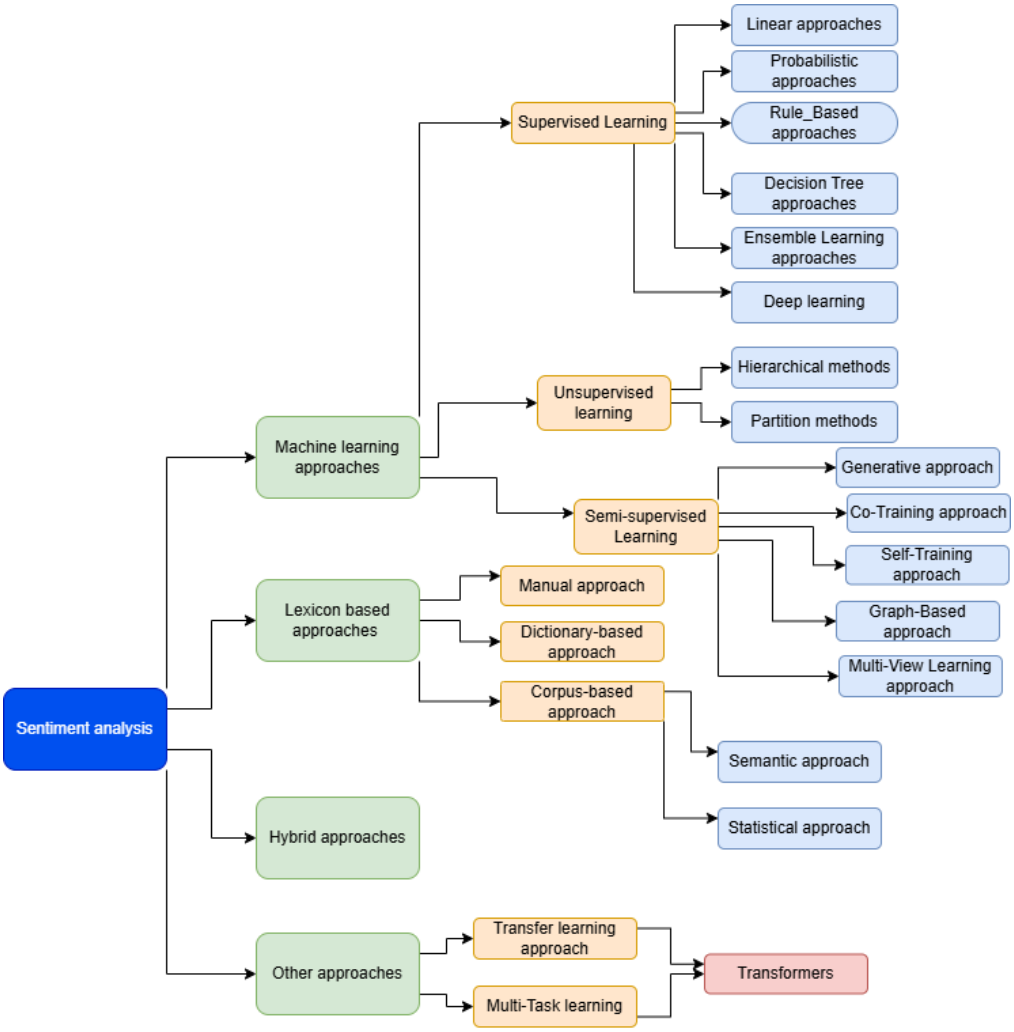



**Figure 2.1:** Overview of sentiment analysis approaches.

### Lexicon-Based approach

The lexicon-based approach relies on predefined sentiment lexicons, which contain words or phrases annotated with their sentiment polarity (i.e., positive, negative, or neutral). This method estimates the overall sentiment of a text by aggregating the sentiment scores of its individual words [13].

Lexicon-based approaches can be categorized into:

- **Dictionary-based methods**: The dictionary-based approach to sentiment analysis relies on the identification of opinion-bearing words through lexical resources. This method begins by manually constructing a seed list of sentiment-laden terms, a process often constrained by the limited availability of domain-specific opinion words. To expand this lexicon, external linguistic resources such as dictionaries and thesauri are consulted to retrieve semantically related synonyms and antonyms. These related terms are iteratively incorporated into the original seed list, allowing the sentiment lexicon to evolve and achieve greater lexical coverage. The process continues until a sufficiently

comprehensive and contextually appropriate set of sentiment expressions is developed, enabling more accurate sentiment classification across varied textual data [14].

- **Corpus-based methods**: The corpus-based approach to sentiment analysis involves the construction and iterative expansion of a list of opinion seed words by leveraging patterns derived directly from domain-specific textual corpora. These seed words typically consist of general sentiment-bearing terms that are contextually grounded in the subject matter of the discourse. Through co-occurrence statistics and contextual association measures, the lexicon is enriched with additional sentiment expressions identified within the corpus. This method often integrates statistical techniques such as pointwise mutual information or chi-square tests with semantic orientation analysis to capture nuanced sentiment variations across different contexts. The corpus-based approach thus facilitates a more adaptable and domain-sensitive sentiment classification, as it reflects the linguistic characteristics and sentiment trends present within the target corpus [14].

Although lexicon-based methods are interpretable and do not require labeled data, they struggle with issues like domain dependency, context sensitivity, and negation handling [15]. For Arabic sentiment analysis, lexicon-based methods often fail to account for dialectal variations, requiring the development of specialized resources such as the Arabic Sentiment Lexicon (ArSenL) [16].

### Machine learning approach

The machine learning approach to sentiment analysis encompasses supervised, semi-supervised, and unsupervised learning paradigms. It involves the automated extraction and utilization of features from textual data to perform classification through a multi-stage analytical process. This methodology is particularly favored in large-scale text classification tasks due to its capacity to process extensive datasets with minimal human intervention. Each of the three approaches—supervised, unsupervised, and semi-supervised—offers distinct advantages depending on the availability of annotated data and the specific application context.

- **Supervised Learning Approach**: Supervised machine learning remains a predominant approach in sentiment analysis classification, widely adopted due to its effectiveness in structured prediction tasks. This method requires two distinct datasets: a labeled dataset used for training the model and a separate dataset employed for evaluation and validation. The performance and reliability of the resulting classifier are highly contingent on the quality and correctness of the annotated training data; mislabeling within the training set can significantly degrade model accuracy. Common algorithms used in supervised sentiment classification include Decision Trees (DT), Naïve Bayes (NB), and Support Vector Machines (SVM), which have been shown to deliver competitive results across various domains [17].

- **Unsupervised Learning Approach**: Unsupervised learning techniques operate without reliance on labeled datasets, making them particularly suitable in scenarios where data annotation is impractical or unavailable. These methods derive patterns and insights directly from input data by identifying inherent structures, often through clustering or dimensionality reduction. Algorithms such as K-means clustering and Word2Vec have been widely utilized in large-scale sentiment analysis, especially in social media contexts, where vast amounts of unstructured data are prevalent. Despite their scalability, unsupervised models necessitate substantial data repositories to achieve meaningful and accurate representations. Furthermore, the absence of labeled guidance can increase the risk of generating ambiguous or inaccurate outputs, potentially necessitating the adoption of more robust supervised learning alternatives [18].

- **Semi-Supervised Learning Approach**: Semi-supervised learning integrates the strengths of both supervised and unsupervised learning paradigms to address limitations associated with the scarcity of labeled data. This approach is particularly advantageous when acquiring labeled examples is resource-intensive, yet a large volume of unlabeled data is available. In semi-supervised frameworks, models are trained on a small subset of labeled data alongside a significantly larger pool of unlabeled data, allowing for improved generalization and performance [19]. By leveraging the structure and distribution of unlabeled data, semi-supervised methods effectively enhance learning accuracy while reducing annotation costs. This approach has shown considerable utility in natural language processing tasks, including sentiment analysis, where annotated corpora are often limited [20].

### Hybrid approach

Hybrid approaches combine lexicon-based and machine-learning techniques to improve sentiment classification accuracy. These methods typically integrate lexicon-based sentiment scores as additional features in machine learning or deep learning models [21]. By leveraging both linguistic knowledge and statistical learning, hybrid models can better handle context-dependent sentiment and nuanced language usage. Some hybrid techniques also utilize ensemble learning, where multiple classifiers are combined to enhance robustness and generalization [22]. This approach helps mitigate individual model weaknesses, leading to improved performance across diverse datasets. Hybrid sentiment analysis methods are particularly useful for languages with complex morphology, such as Arabic, where context greatly influences meaning.

### Transfer Learning Approach to Sentiment Analysis

The widespread and continuous use of social media for cross-platform communication has generated vast volumes of user-generated content, necessitating automated sentiment analysis systems capable of domain-specific generalisation. Traditional

deep learning approaches, such as the convolutional neural networks (CNNs) employed by Sharath and Tandon [23], have been utilized for tweet-based sentiment classification. Although these models demonstrated efficacy, they required extensive domain-specific training corpora, which limits scalability and adaptability across domains and dialects.

To address such limitations, transfer learning has emerged as a robust alternative, enabling pretrained language models to be fine-tuned on specific sentiment analysis tasks. Rather than training models from scratch, transfer learning leverages generalised linguistic knowledge from large-scale pretrained models and adapts this knowledge often in low-resource or domain-specific contexts. This is particularly advantageous for Arabic sentiment analysis, where the variability in dialects, slang, and domain-specific expressions complicates traditional model training.

Authors in [24] and [25] highlighted challenges arising from informal text, emoticons, and contextual ambiguity in Twitter data. Deep learning models that integrate transfer learning, such as fine-tuned transformers, offer improved performance in such cases by encoding semantic and syntactic patterns during pretraining. These models—e.g., BERT or AraBERT [26] can retain sequential and contextual information while being adapted to sentiment classification tasks through supervised fine-tuning.

Unlike conventional CNNs, which require extensive domain-specific data and manual feature engineering, transfer learning models capture deep linguistic representations that generalise across topics and text types. [27] advocates for neural language models capable of generating word embeddings that reflect complex semantic structures. These capabilities make transfer learning particularly effective in multilingual or morphologically rich contexts languages, such as Arabic.

### Transformers

Transformers, introduced by Vaswani in [28], represent a seminal advancement in the field of natural language processing (NLP), providing a framework that significantly departs from earlier sequential models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs). Unlike its predecessors, which process input sequences sequentially in a step-wise manner, the Transformer architecture employs a fully attention-based mechanism. This design enables parallelization during both training and inference, resulting in significantly improved computational efficiency and scalability. At the core of the Transformer model, as illustrated in Figure 2.2, lies the self-attention mechanism, which allows the model to weigh the relevance of different tokens within a sequence when constructing contextual representations.

**self-attention mechanism** enables the network to weigh the importance of different tokens in an input sequence relative to one another, regardless of their positional distance. This is accomplished through the computation of scaled dot-product attention, where the input is transformed into queries, keys, and values, allowing the model to dynamically adjust its focus based on the contextual relevance of tokens. The architecture of transformer consists of an encoder-decoder structure, with each encoder and decoder layer composed of multi-head self-attention

mechanisms and position-wise feed-forward networks. Additionally, the use of positional encoding compensates for the absence of recurrence, encoding the order of the sequence directly into the input embeddings.

**Transformer** are particularly well-suited for transfer learning due to their capacity to model contextual relationships in text through self-attention mechanisms and their ability to scale effectively with increased data and computational resources [28]. Pretrained Transformer-based models, such as BERT [29], serve as foundational encoders or decoders that can be fine-tuned on a wide range of tasks, including sentiment analysis, question answering, and named entity recognition.

From a theoretical standpoint, the Transformer's ability to model long-range dependencies and capture global context without suffering from vanishing gradients has been critical to its success. Furthermore, its design facilitates fine-tuning on downstream tasks through transfer learning, which has become a prevalent paradigm in modern NLP research and applications.

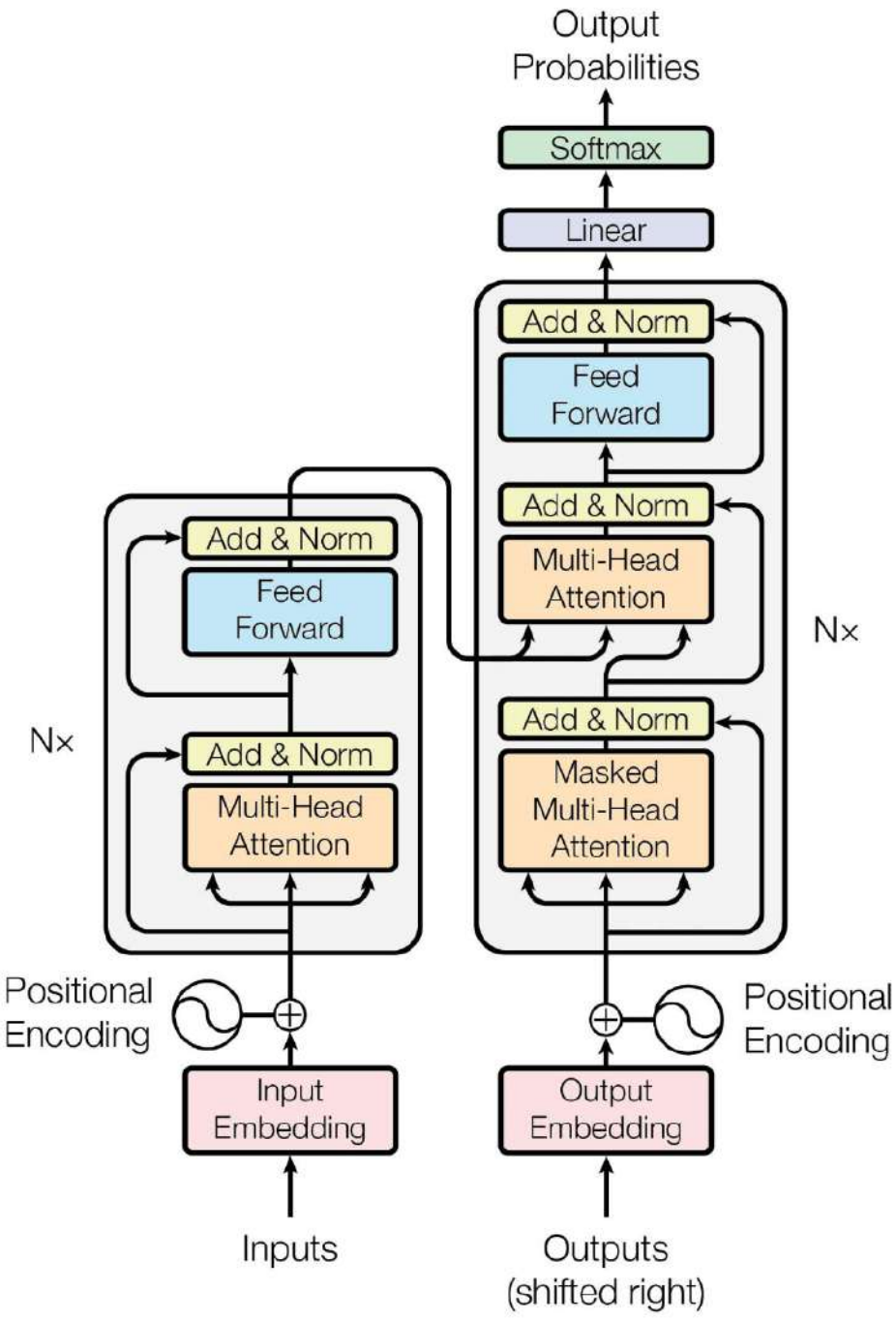



**Figure 2.2:** The encoder-decoder structure of the transformer architecture [28].

### 2.2.2 Sentiment analysis challenges

Despite the significant advancements, sentiment analysis still faces several challenges including:

- **Ambiguity and Sarcasm**: Many sentiment expressions exhibit ambiguity

or sarcasm, making it difficult for models to accurately determine sentiment polarity. Sarcasm often conveys negative sentiment using positive words, which can mislead traditional classifiers. For example, the phrase "Great! Another delay!" appears positive but expresses frustration. Addressing such challenges requires context-aware models and advanced linguistic understanding [30].

- **Domain-Specificity**: The meaning of sentiment-laden words varies across domains, necessitating domain adaptation techniques. For instance, the word "hot" is positive in the context of food (e.g., "hot and delicious soup") but negative in weather-related discussions (e.g., "hot and unbearable day"). Sentiment analysis models trained on generic datasets often struggle when applied to domain-specific data [31].

- **Negation Handling**: The presence of negation words (e.g., not, never, hardly) significantly alters sentiment meaning. For example, "not bad" carries a positive sentiment despite containing a negative word. Traditional sentiment analysis approaches often fail to capture such nuanced expressions, requiring linguistically sophisticated models that incorporate negation scope detection and context-aware sentiment reversal techniques [32].

- **Data Imbalance**: Many sentiment analysis datasets suffer from class imbalance, where the neutral sentiment class is overrepresented compared to positive and negative sentiments. This imbalance can lead to biased models that favor the dominant class, reducing classification performance for minority sentiment categories. Re-sampling techniques, cost-sensitive learning, and data augmentation methods are often employed to mitigate this issue and improve model robustness [33].

- **Linguistic Diversity Across Languages** Sentiment analysis systems often struggle with the wide variability in linguistic structures among different languages. Each language exhibits unique syntactic, semantic, and grammatical rules, making it difficult to design universal models that perform consistently across multilingual settings [34].

## 2.3 Arabic language in natural language processing

Arabic is a widely spoken Semitic language with unique linguistic characteristics that present both challenges and opportunities in NLP. Its complex morphology, dialectal diversity, and script ambiguity make it a critical research area. Arabic consists of Modern Standard Arabic (MSA) and numerous dialects, each with distinct grammatical and lexical variations, complicating NLP tasks like tokenization, sentiment analysis, and machine translation.

Arabic Natural Language Processing (ANLP) has enjoyed increasing attention in recent years, with several cutting-edge systems developed for a wide range of applications, including sentiment analysis, sarcasm detection, speech synthesis,

machine translation, data retrieval, and text-to-speech conversion [35]. These systems face a range of complex issues due to the unique structure of Arabic, necessitating specialized models capable of handling these challenges. ANLP applications must address several inherent linguistic features of the Arabic language, such as its rich morphology and diverse dialects. Consequently, successful systems must be capable of reconciling these linguistic complexities to provide effective and consistent language processing outcomes.

This section examines key challenges, characteristics, and features of Arabic NLP, providing examples of its practical applications.

### 2.3.1 Difficulties with Arabic language

Although natural language processing (NLP) has made significant progress, Arabic continues to pose considerable challenges due to its complex linguistic and structural features:

#### Morphological Complexity

Arabic is a morphologically rich language, employing a root-and-pattern system where words are derived from three-letter roots combined with specific patterns to convey different meanings. This nonconcatenative morphology leads to a high degree of inflection and derivation, complicating tasks such as tokenization, lemmatization, and part-of-speech tagging. For instance, the root "k-t-b" كتب can generate words like "kataba" (he wrote), "maktab" (office), and "maktaba" (library) [36].

#### Diglossia

Arabic exhibits diglossia, characterized by the coexistence of Modern Standard Arabic (MSA) and various regional dialects. MSA is used in formal contexts, while dialects are prevalent in everyday communication. The significant differences between MSA and dialects pose challenges for NLP systems, as models trained on MSA may not perform well on dialectal text [36].

#### Dialectal Variability

Arabic dialects vary significantly across regions, affecting vocabulary, pronunciation, and syntax. This variability means that NLP models trained on one dialect may not generalize well to others, necessitating the development of multi-dialectal resources and models [37].

#### Orthographic Challenges

The Arabic script is written from right to left, and letters change shape based on their position in a word. Additionally, short vowels (diacritics) are often omitted

in writing, leading to ambiguity. For example, the sequence "ktb" كتب can be read as "kataba" كَتَبَ (he wrote) or "kotob" كُتُبْ (books), depending on the context [36].

## 2.4 Characteristics of the Arabic language

### 2.4.1 The Arabic Language Features

The Arabic language, a Semitic language known for its complex and unique morphology, is the official language of 22 countries and spoken by over 400 million people globally [38]. It ranks as the fourth most widely used language on the internet and is among the top ten most-used languages online [39]. Arabic uses an alphabet of 28 letters and is written from right to left. Its structure follows a free word order governed by specific rules, with a morphology based on the derivation of root words through various intonations [40].

Arabic exists in two primary forms: Modern Standard Arabic (MSA), the formal and standardized variant used in media, literature, and academic settings across the Arab world, and dialectal Arabic, the colloquial form used in everyday communications. MSA adheres to classical grammatical structures derived from the Quran and boasts a vast vocabulary. In contrast, dialectal Arabic, while rooted in MSA, varies significantly across regions, with distinct lexical choices, grammar, and phonology specific to different countries [41]. Social media communication predominantly employs dialectal Arabic, which diverges from MSA in terms of phonology, morphology, and syntax.

Dialectal Arabic is divided into several major subgroups, including Levantine, Gulf, Maghrebi, Egyptian, Iraqi and others [36]. This linguistic diversity, along with the absence of standardized references, poses significant challenges for ANLP. ANLP systems must account for the contextual variations in meaning and structure that arise from these different forms of the language. For example, the root word "كَتَبْ" (katab, "write") can generate various forms, such as "تكْتُبْ" (taktob, "she writes") and "كَتَبَتْ" (katabat, "she wrote"), which differs from how language processing works in English and other languages.

### 2.4.2 The Difference Between English and Arabic Language

Significant linguistic divergences exist between English and Arabic, encompassing lexical, grammatical, and syntactic dimensions. Lexical discrepancies include phenomena such as non-vocalisation, lexical insufficiencies, polysemy, and variances in connotation and collocation. On the grammatical and syntactic levels, contrasts emerge from differences in word order, gender and referential structure, verb tense and aspect usage, prepositional structures, definite article usage, and coordination mechanisms. Unlike English, which includes silent letters and complex orthographic combinations, Arabic follows a largely phonetic spelling system in which each letter

corresponds to a distinct phoneme. For instance, in the word ‘thing’ the ‘th’ sound in the English language is reduced to the ث character in Arabic .

Arabic further differs in that it does not utilize a linking verb equivalent to “to be” in present tense constructions and lacks an indefinite article altogether. Moreover, Arabic script is inherently case-insensitive and is written from right to left. This results in significant morphological variation, as the shape of a letter changes depending on its position within a word—initial, medial, or final. Letters with connective properties can be joined in both handwritten and typographic contexts.

Gender in Arabic is expressed grammatically and naturally: all nouns are either masculine or feminine. Natural gender pertains to animate beings, whereas grammatical gender applies to inanimate objects. Feminine nouns are often formed by appending the suffix "ت" (ta) or "ة" (ta marbuta) to the masculine form [42] 2.4.3. These fundamental structural differences render the direct application of English-based Natural Language Processing (NLP) tools to Arabic inadequate. Consequently, dedicated ANLP frameworks are required to address the unique linguistic features of the Arabic language.

### 2.4.3 Examples of Arabic Natural Language Processing

#### Singular and Plural

In both Modern Standard Arabic (MSA) and dialectal Arabic, non-imperative sentences exhibit morphological variation in verb forms depending on the gender of the referent. These gender-based verbal inflections are systematically reflected in verb morphology [43], as illustrated in Tables 2.1, 2.2 and 2.3.

**Table 2.1:** The Arabic Language Features - Sentences phrased by male and female speakers in singular examples.

| English | Arabic | Gender |
|---|---|---|
| He wrote the report | هو كتب التقرير<br>howa kataba altaqreer | Masculine |
| She wrote the report | هي كتبت التقرير<br>hiyya katabat altaqreer | Feminine |

Sentences phrased by male and female speakers in different tenses.

**Table 2.2:** The Arabic Language Features - Sentences phrased by male and female speakers in different tenses.

| English | Arabic | Tense | Gender |
|---|---|---|---|
| I write the report | أكتب التقرير<br>akatob altaqreer | Present simple | Masculine / feminine |
| I wrote the report | كتبت التقرير<br>katabat altaqreer | Past simple | Masculine / feminine |

Sentences phrased by male and female speakers in plural.

**Table 2.3:** The Arabic Language Features - Sentences phrased by male and female speakers in plural examples.

| English | Arabic | Tense | Gender |
|---|---|---|---|
| We write the report | نكتب التقرير<br>naktob altaqreer | Present simple | Masculine / feminine |
| We wrote the report | كتبنا التقرير<br>katabna altaqreer | Past simple | Masculine / feminine |

### Gender

Nouns may be assigned to gender classes in languages that exhibit grammatical gender systems [44]. In such systems, the gender of a noun is often determined by its morphological or phonological structure. This feature presents challenges when translating grammatical gender from English into Arabic—both in Modern Standard Arabic (MSA) and dialectal varieties—due to structural and semantic differences. The literature highlights that gender-related ambiguities frequently arise from generalisations in English pronouns such as "I," which are rendered as أنا in Arabic. The accurate treatment of gender distinctions is particularly critical in Arabic due to its sensitivity to gender agreement in verbs, adjectives, and pronouns [45]. Illustrative examples are provided in Tables 2.4,2.5,2.6 and 2.7.

**Table 2.4:** The Arabic Language Features – Gender examples.

| English | Arabic | Gender |
|---|---|---|
| I am a tourist (male) | أنا سائح<br>ana sae'h | Masculine |
| I am a tourist (female) | أنا سائحة<br>ana sae'hah | Feminine |

The concept of grammatical gender comprises a two-tiered semantic framework. At its core, it often aligns with biological distinctions, denoting male and female categories. In Arabic, this is exemplified by gender-specific nouns such as رجل (rajul, man) and امرأة (imra'ah, woman). However, in contrast, certain occupational or role-based nouns, such as "doctor" or "driver," are generally treated as gender-neutral unless contextually specified. Arabic morphology incorporates gender markers, and many nouns exhibit gender-specific morphological features. Typically, nouns ending in characters such as ى (alif maqsura), ة / ه (taa marbuta), or اء (alif hamza) are classified as feminine [44]. These morphological cues serve as essential indicators for gender agreement in Arabic syntax and semantics.

### Suffixing

### Suffixing – The linked taa'(ـه)

**Table 2.5:** The Arabic Language Features - Suffixing – The linked taa' examples.

| English | Arabic (male) | Arabic (female) |
|---|---|---|
| Lawyer | محامي<br>mohami | محاميه<br>mohamiyah |
| Secretary | سكرتير<br>secretaire | سكرتيره<br>secretairah |

### Suffixing – al Alif al Maqsūra (ى)

**Table 2.6:** The Arabic Language Features - Suffixing – al Alif al Maqsūra examples.

| English | Arabic | Gender |
|---|---|---|
| The oldest brother | الأخ الأكبر<br>al'akh al'akbar | Masculine |
| The oldest sister | الأخت الكبرى<br>al'okht alkobra | Feminine |

### Suffixing – al Alif al Mamdūdah (ـاء)

**Table 2.7:** The Arabic Language Features - Suffixing – al Alif al Mamdūdah examples

| Single | Arabic (male) | Arabic (female) |
|---|---|---|
| Single | أعزب<br>a'azab | عزباء<br>'azbaa' |

### Negation

Negation in English is expressed through the use of auxiliary verbs and particles such as "do not," "does not," "did not," and "no," which function to reverse the polarity of a statement. Examples illustrating these forms are presented in Table 2.8.

**Table 2.8:** The Arabic Language Features – Negation examples.

| | English | Arabic (MSA and dialectal Arabic) |
|---|---|---|
| Do not, does not = لا | Do not give up | لاتستسلم<br>la tastaslem |
| Did not = لم | He did not sleep all night | لم ينم طوال الليل<br>lam yanam twal allayl |
| No= ليس / لا / ممنوع | smoking No | ممنوع التدخين<br>mamnoo‘ altadkheen |

### Free Word Order

Arabic is characterized by a relatively free word order, a feature that significantly distinguishes it from English. According to [46], the Arabic language accommodates several permissible syntactic structures, including subject-verb-object (SVO), verb-subject-object (VSO), verb-object-subject (VOS), and object-verb-subject (OVS) [47]. This syntactic flexibility enriches Arabic's grammatical complexity and makes it uniquely suited for various expressive constructions. In contrast, English predominantly adheres to the SVO structure, and deviations from this norm are typically ungrammatical. Table 2.9 illustrate these differences, where all configurations are considered syntactically valid in Arabic (marked with checkmarks), while only the SVO order is acceptable in English, with other arrangements denoted as incorrect (xmarks).

**Table 2.9:** The Arabic Language Features - Free word order examples.

| Sentence Form | English | Arabic |
|---|---|---|
| SVO | The girl walks slowly ✓ | البنت تمشي ببطء✓<br>albento tamshi bebote’ |
| VSO | Walks the girl slowly ✗ | تمشي البنت ببطء✓<br>tamshi albento bebote’ |
| VOS | Walks slowly the girl ✗ | تمشي ببطء البنت✓<br>tamshi bebote’ albento |
| OVS | Slowly walks the girl ✗ | ببطء تمشي البنت✓<br>bebote’ tamshi albento |

Arabic exhibits morphological agreement between the subject and verb, allowing

for considerable syntactic flexibility through free word order. This linguistic feature enables the distinction between the subject and object based on inflectional markings rather than fixed word positions. The semantic equivalence of various word orders in Arabic stands in contrast to English, which adheres to a rigid subject-verb-object (SVO) structure [48]. For instance, while adverbs in English typically precede the verbs they modify, Arabic permits adverbs to appear either before or after the verb. Consequently, the morphological and syntactic analysis of Arabic is substantially more complex than in English and other Indo-European languages, particularly due to its rich inflectional morphology and syntactic variability.

### Proper Nouns in the Arabic Language

A closer examination of gender shifts in referents reveals that such changes have significant implications at both the phrasal and sentential levels. These morphological variations influence adjective conjugation, agreement, and syntactic structure, as demonstrated in Table 2.11.

**Table 2.10:** The Arabic Language Features - Proper Nouns in the Arabic Language examples.

| English | | Arabic |
|---|---|---|
| A big castle | In Arabic, castle = masculine,<br>which makes big = masculine | قصر كبير<br>qaser kabeer |
| A big car | In Arabic, car = feminine,<br>which makes big = feminine | سياره كبيره<br>sayarah kabeerah |

### Changing Verbs According to Gender

Authors in [49] have emphasized that in Arabic, the verb is influenced by the gender of the subject. This variation extends beyond the verb itself, affecting multiple lexical components within a sentence. Table 2.10 illustrates these gender-dependent morphological and syntactic shifts.

**Table 2.11:** The Arabic Language Features - Changing Verbs According to Gender examples.

| English | Arabic | Gender |
|---|---|---|
| One of my students did not attend the session | لم يحضر احد طلابي الجلسة<br>lam yahdor ahad tolabbi algalsah | Masculine |
| One of my students did not attend the session | لم تحضر احدى طالباتي الجلسه<br>lam tahdor ehda talebati algalsah | Feminine |

**Using a Nominal Phrase with the Pronouns 'He' or 'She'**

Authors in [50] notes that in English, noun phrases such as "a smart manager" are gender-neutral; neither the determiner nor the adjective requires inflection to match the gender of the head noun. In contrast, Arabic exhibits full agreement in gender and number between the determiner, adjective, and the head noun. Consequently, when the head noun is feminine, both the adjective and determiner must reflect this morphological change. This distinction is exemplified in Table 2.12.

**Table 2.12:** The Arabic Language Features - Using a Nominal Phrase with the Pronouns 'He' or 'She' examples.

| English | Arabic | Gender |
|---|---|---|
| He is a smart manager | هو مدير ذكي<br>howa modeer dhaki | Masculine |
| She is a smart manager | هي مديرة ذكية<br>hiyya modeerah dhakiyah | Feminine |

### 2.4.4 Morphological differences between MSA and Arabic dialect

Modern Standard Arabic (MSA) and Arabic dialects differ significantly in morphology, with crucial implications for natural language processing applications. MSA exhibits a templatic morphological system rooted in classical Arabic, characterized by root-pattern structures and extensive affixation [51]. For example, the verb kataba (كَتَبَ, "he wrote") derives from the root k-t-b and appears in templatic forms like yaktubu يَكْتُبْ ("he writes") and maktūb مَكْتُوبْ("written"), marking voice, tense, and case through inflectional morphology.

Conversely, Arabic dialects display a simplified morphological systems that diverge from MSA. Dialects often eliminate categories such as dual number and grammatical case. For example, the MSA dual form kitaban (كِتَابَانِ, "two books") is rendered as kitabin (كِتَابَيْنِ) in Egyptian Arabic, merging the dual with the plural [52]. Dialectal verb conjugations also diverge; the MSA future marker (س- or سَوْفَ) becomes ha (ها) in Egyptian Arabic (e.g., هاكَلِّمْكِ, "I will call you") and Rah (

رَح)- in Levantine (e.g., رَح اِحْكِي مَعَكَ, "I will talk to you").

Morphological innovation in dialects is further shaped by contact with colonial languages. For instance, Maghrebi Arabic incorporates French morphology, evident in forms like telefonīt (تيليفونيت) ("I phoned"), adapted from téléphoner [52]. Dialect-specific clitics also vary significantly: the MSA object pronoun -hu (ـهُ, "him") becomes -o or -u in dialects, as in shafto (شَافِتو) ("she saw him") in Egyptian Arabic [53].

These variations complicate core NLP tasks such as sentiment analysis, hate speech detection, or sarcasm detection.

### 2.4.5 Characteristics of the Arabic Language Relevant to Sentiment Analysis

Authors in [54] emphasize that Arabic is a morphologically rich language, wherein extensive syntactic and semantic information is encoded at the word level. Unlike English, which exhibits relatively limited morphological variation, Arabic words can manifest in numerous surface forms due to their complex inflectional morphology. This linguistic richness presents significant challenges in adapting sentiment analysis systems originally developed for English, particularly those that rely on lexical-level features. Directly applying such systems to Arabic often results in data sparsity and reduced accuracy, as a single Arabic lemma may correspond to multiple inflected forms, as illustrated in Table 2.13.

**Table 2.13:** Example of multiple forms of Arabic verbs .

| English word | Arabic word | Forms in English | Forms in Arabic |
|---|---|---|---|
| love | حب (root) hub | I love | أُحِبُّ uhibo |
| | | He loves | يُحِبُّ yuhibu |
| | | She loves | تُحِبُّ tuhibu |
| | | They love | يُحِبُّونَ yuhbuna |
| | | We love | نُحِبُّ nahibu |

In English, verbs such as love exhibit limited morphological variation and typically maintain a consistent lexical identity. Conversely, in Arabic, a single

verb root may generate a wide array of surface forms due to its morphologically rich structure. This phenomenon significantly complicates sentiment analysis, as the same lexical root may produce dozens of derivational or inflectional variants. Compounding this issue, many Arabic first names, especially family names, originate from adjectives, thereby increasing the potential for lexical ambiguity in sentiment classification [55]. As illustrated in Table 2.14, this overlap between proper nouns and sentiment-bearing adjectives poses a unique challenge in Arabic NLP. Traditional solutions employ part-of-speech (POS) tagging to distinguish between proper nouns and adjectives through pattern recognition. However, this approach is notably less effective in dialectal Arabic due to the reduced accuracy of available POS taggers for non-standardized varieties.

**Table 2.14:** Examples of Arabic names.

| Arabic name | Adjective |
|---|---|
| نبيل Nabil | النبل Noble |
| سعيد Saeid | السعادة Happy |
| جميله Jamiluh | الجمال Beautiful |

The presence of diacritics and the morphologically rich nature of the Arabic language result in multiple words derived from the same root that may convey divergent or even contradictory emotional orientations. This complexity presents a significant obstacle in sentiment analysis, particularly when employing stemming techniques intended to reduce words to their roots for polarity identification. Such approaches may lead to incorrect sentiment classification, as semantically incompatible words can share identical roots. Table 2.15 illustrates examples of sentimentally inconsistent terms that originate from the same Arabic root.

**Table 2.15:** Arabic is morphologically rich.

| Arabic word | In English | Sentiment | Root |
|---|---|---|---|
| تلاعب talaeub | Manipulate | Negative -1 | لعب laeib |
| يلعب yaleab | Plays | Positive +1 | |
| تمييز tamyiz | Discrimination | Negative -1 | ميز miz |
| إمتياز iimtiaz | Excellent | Positive +1 | |

The challenges outlined in this section underscore the necessity of applying advanced linguistic preprocessing techniques for effective sentiment analysis of Arabic tweets. A thorough understanding of the linguistic features of Arabic and the complexities inherent in social media content highlights that Natural Language Processing (NLP) tools developed for Modern Standard Arabic (MSA)

often exhibit limited efficacy when applied to dialectal Arabic. Furthermore, dialectal variations differ significantly across Arabic-speaking regions; hence, NLP tools tailored for one dialect, such as Egyptian Arabic, may perform poorly when applied to others, such as Saudi Arabic.

### 2.4.6 Arabic transformers

Transformers have significantly advanced NLP, particularly in tasks involving low-resource languages such as Arabic. Given the complexity of the language, transformer-based models trained specifically on Arabic data have become essential for achieving state-of-the-art performance in various NLP tasks. Early attempts to apply multilingual models, such as mBERT and XLM-RoBERTa, revealed promising results; however, their performance on Arabic-specific tasks was limited due to the underrepresentation of Arabic data during pretraining [39, 56].

To address these challenges, several Arabic-specific transformer models have been developed. AraBERT, introduced by Antoun et al. [57], was one of the first monolingual BERT-based models trained on large-scale Arabic corpora, including news and social media text. AraBERT significantly improved the accuracy of downstream tasks such as sentiment analysis and named entity recognition. Subsequent models like AraELECTRA [58] and MARBERT [26] extended this work by incorporating additional dialectal data and leveraging different pretraining objectives. For instance, MARBERT was specifically trained on a massive corpus of dialectal Arabic tweets, demonstrating superior performance in sentiment analysis and dialect identification.

These models contribute to a growing ecosystem of Arabic NLP resources, enabling fine-tuned applications in areas like hate speech detection, sarcasm identification, and multi-dialect classification. They also highlight the importance of pretraining on diverse and representative data to ensure robustness across dialects and domains.

## 2.5 Multi-Task Learning for NLP

Humans possess the remarkable ability to learn multiple tasks simultaneously, often transferring knowledge acquired from one task to facilitate learning in another. For instance, the motor skills and strategies developed while learning to play tennis can enhance the acquisition of squash skills, and vice versa. This cognitive capability has inspired the development of Multi-Task Learning (MTL) in machine learning, a strategy designed to emulate such human learning patterns. MTL seeks to jointly train models on multiple related tasks, enabling the sharing of inductive biases and learned representations across tasks. By leveraging inter-task relationships, MTL aims to improve the generalization performance of all tasks involved, particularly when they exhibit underlying commonalities or share domain-specific features [59].

### 2.5.1 Foundations of Multi-Task Learning

A fundamental prerequisite for effective multi-task learning (MTL) is the relatedness between tasks and their associated data. MTL is most effective when tasks are positively correlated, that is when they share similar objectives or overlapping data distributions. These tasks mutually benefit from shared lower-level representations, enhancing prediction consistency across tasks [59]. In MTL frameworks, the hidden representations learned by one task are often preferred by others, which allows feature sharing and fosters the discovery of complex feature interactions.

Recent advances in deep neural architectures, such as BERT [29], have facilitated MTL through flexible encoder-decoder structures capable of adapting to various tasks with minimal modification. As authors in [60] demonstrated, larger networks tend to yield better performance in MTL settings due to increased capacity for learning task-specific and task-invariant features.

MTL enhances data efficiency by enabling each task to learn from the information encoded in related tasks. Moreover, tasks often exhibit distinct noise distributions, and when trained together, this diversity acts as an implicit data augmentation mechanism, encouraging more robust and generalizable feature representations. This mitigates overfitting and improves performance on the related tasks [61]. MTL is particularly advantageous for low-resource tasks, which can be improved through co-training with high-resource tasks from related domains [62] [63] thereby amplifying training signals and stabilizing learning outcomes.

Furthermore, recent studies have shown that multi-task models often converge faster than their single-task counterparts, as auxiliary tasks can provide gradient signals that help escape poor local minima and guide optimization towards more effective regions of the parameter space [64]. This leads not only to improved predictive performance but also to more stable and interpretable learning outcomes.

### 2.5.2 Components of MTL for Text Classification

To implement MTL in text classification, different components are designed to balance shared feature learning and task-specific customization. These components must be optimized to prevent negative transfer (where one task detracts from the performance of another). Below are key architectural strategies employed in MTL for text classification:

#### Shared and Task-Specific Layers

Shared and task-specific layer architectures represent one of the most fundamental and widely adopted designs in multi-task learning (MTL), particularly in text classification. As shown in 2.3 This architecture segregates the model into two major components: shared layers and task-specific layers. Shared layers, typically comprising embeddings, recurrent or Transformer blocks, whereas Task-specific layers appended atop the shared structure, are responsible for learning discriminative features tailored to each task's objective [64].

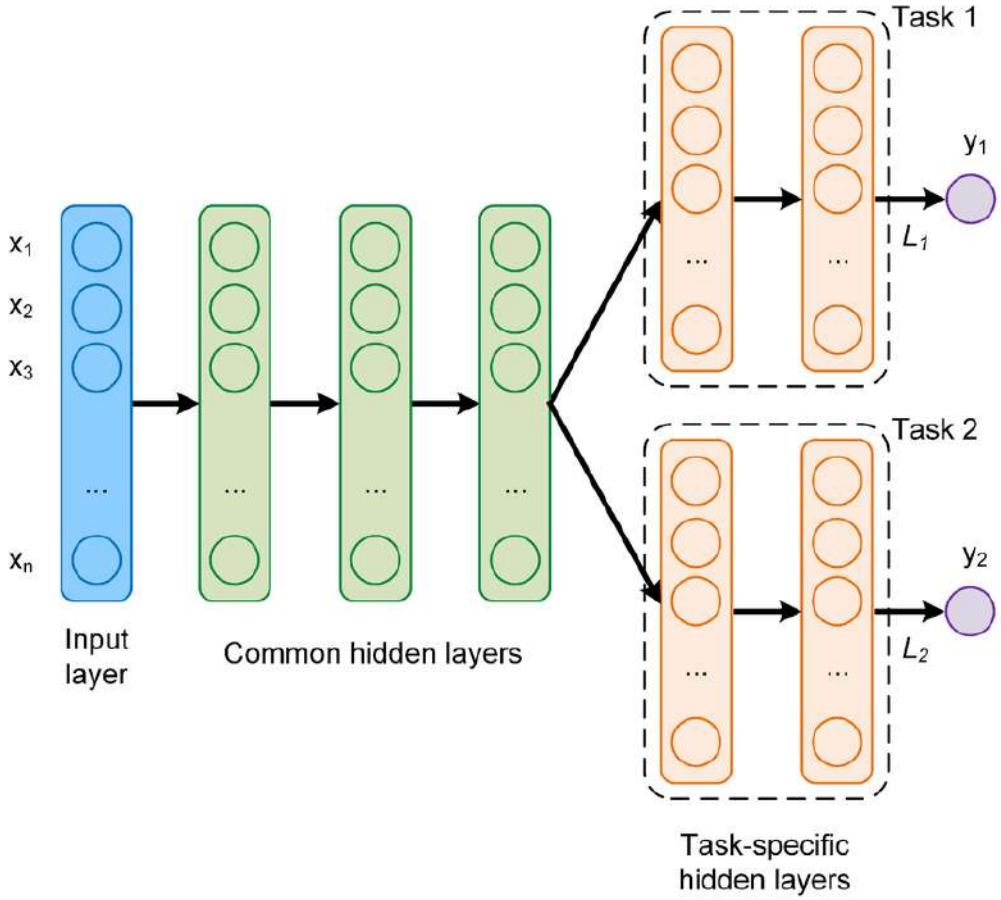



**Figure 2.3:** MTL Shared and Task-Specific Layers architecture.

### Encoder-Decoder architectures

Encoder-decoder architectures have also been explored in multi-task learning (MTL) frameworks for sentiment analysis, particularly when the problem is framed as a text-to-text task. In such settings, the architecture as illustrated in Figure 2.4 comprises a shared encoder that transforms input text into a high-dimensional latent representation, followed by task-specific decoders that generate outputs tailored to each downstream task. While encoder-decoder models are traditionally more prevalent in generative tasks such as text summarization or question answering, recent work has adapted them for classification-oriented tasks like sentiment analysis by rephrasing classification as a generation problem. This design allows the model to leverage shared semantic representations across tasks while accommodating task-specific output spaces, thereby improving generalization and robustness in low-resource or multi-dialect scenarios [65] [66].

The encoder-decoder paradigm in MTL enables both information sharing and task specialization. By leveraging a shared encoder, the model captures universal linguistic and contextual features across tasks, while the decoders adapt these representations for specific outputs [67].

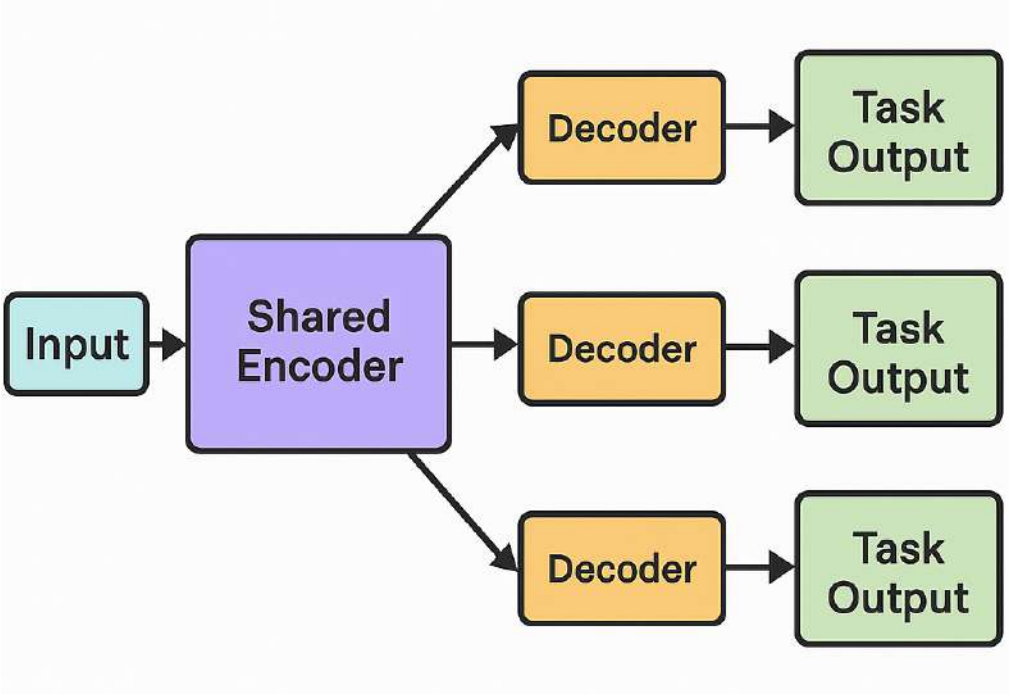



**Figure 2.4:** MTL Encoder-Decoder architectures.

### Parameter Sharing Strategies

In MTL parameter sharing strategies allows for interaction and information flow between tasks, facilitating the sharing of learned representations across tasks. MTL parameter sharing has two main approaches 2.5: hard parameter sharing and soft parameter sharing.

**Hard Parameter Sharing** Hard parameter sharing is the most commonly used approach in MTL for text classification [59]. The models in this paradigm share the lower layers (often encoder layers) across all tasks, while the upper layers are for task-specific outputs. This method significantly reduces the risk of overfitting and allows efficient representation learning across tasks. The shared layers capture general linguistic or semantic features, while task-specific layers adapt this representation to individual task requirements.

Models such as the Multi-Task BiLSTM or Multi-Task CNN architectures are frequently used in this context, especially in sentiment analysis and sequence classification tasks [68] [69]. Hard sharing is particularly effective when the tasks are closely related.

**Soft Parameter Sharing** Soft parameter sharing offers more flexibility than hard sharing by maintaining separate models for each task, but with constraints that encourage the parameters to be similar [70]. This architecture allows the model to learn more nuanced representations for each task while still benefiting from cross-task information.

This approach is more suited to tasks that are only loosely related or vary significantly in linguistic structure. For instance, a model designed to jointly perform named entity recognition (NER) and text classification may adopt soft sharing to accommodate the syntactic complexity of NER and the semantic focus of classification [64].

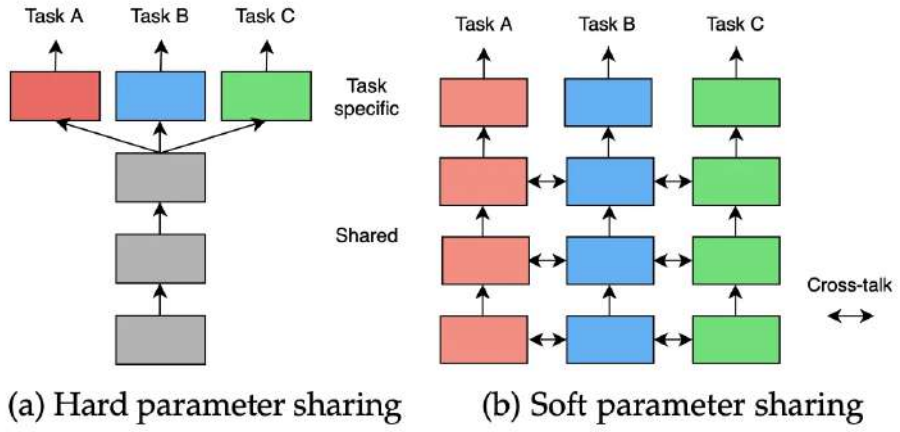



**Figure 2.5:** MTL Hard and Soft Parameter Sharing architectures.

#### Multi-Head Attention for MTL

Multi-head attention has become a fundamental mechanism in deep learning architectures, particularly in Transformer models, due to its ability to capture complex feature interactions across multiple representation subspaces [28]. In the context of multi-task learning (MTL), multi-head attention offers unique advantages by enabling the model to attend differently to shared information for each task. Through parallel attention heads, the model can learn task-specific patterns while simultaneously leveraging shared knowledge across tasks. Recent research has extended multi-head attention for MTL by designing task-specific attention heads, where each head focuses on optimizing a particular task while allowing inter-task communication through shared heads [9] [71]. This structure not only improves task-specific performance but also enhances the generalization of the shared representation. Furthermore, adaptive attention mechanisms have been proposed, wherein the model dynamically selects or reweights attention heads based on the task context, further promoting efficient knowledge sharing and reducing negative transfer [72]. Thus, multi-head attention frameworks significantly contribute to advancing MTL by offering flexible, dynamic, and scalable representations suited to the varying needs of multiple learning objectives.

### 2.5.3 Evaluation and Optimization in Multi-Task Learning

Multi-task learning (MTL) requires specialized optimization strategies to address several challenges: differing task difficulties, unbalanced gradients, and conflicts between task objectives. Without appropriate optimization, negative transfer can

occur, where learning one task degrades performance on others. Effective evaluation and optimization strategies ensure robust, balanced learning across multiple tasks.

The canonical MTL objective is expressed as minimizing a weighted sum of task-specific losses:

$$\mathcal{L}_{\text{MTL}} = \sum_{i=1}^{T} \lambda_i \mathcal{L}_i(\theta) \tag{2.1}$$

where $T$ denotes the number of tasks, $\mathcal{L}_i(\theta)$ is the loss for task $i$ given model parameters $\theta$, and $\lambda_i$ are task-specific weights [73]. Simple static weighting often leads to suboptimal learning, motivating the need for dynamic and adaptive strategies.

### Dynamic Weighting Strategies

Dynamic weighting mechanisms adjust $\lambda_i$ during training to accommodate varying task difficulties. Authors in [74] introduced uncertainty-based weighting, assigning lower importance to tasks with higher homoscedastic uncertainty:

$$\mathcal{L}_{\text{MTL}} = \sum_{i=1}^{T} \left( \frac{1}{2\sigma_i^2} \mathcal{L}_i(\theta) + \log \sigma_i \right) \tag{2.2}$$

where $\sigma_i$ is a learnable parameter representing the uncertainty of task $i$. This method stabilizes training, particularly for tasks combining classification and regression.

Standley et al.[75] proposed Dynamic Weight Averaging (DWA), which adjusts task weights based on the relative rate of loss descent, prioritizing slower-converging tasks to prevent domination by easier tasks.

### Gradient Balancing Techniques

Gradient balancing addresses the disparity in gradient magnitudes across tasks. GradNorm [76] dynamically scales task gradients to match a target training rate:

$$\mathcal{L}_{\text{GradNorm}} = \sum_{i=1}^{T} \left| G_i - \hat{G}_i \right| \tag{2.3}$$

where $G_i = \|\nabla_\theta(\lambda_i \mathcal{L}_i)\|_2$ represents the norm of the task gradient, and $\hat{G}_i$ adjusts according to the relative decrease in task losses.

### Conflict Resolution in Gradients

MTL often suffers from gradient interference, where gradients from different tasks point in opposing directions. To mitigate this, Chen et al. [77] introduced PCGrad (Projected Conflicting Gradient), modifying gradients through projection:

$$\mathbf{g}_i \leftarrow \mathbf{g}_i - \frac{\mathbf{g}_i^\top \mathbf{g}_j}{\|\mathbf{g}_j\|^2} \mathbf{g}_j \tag{2.4}$$

whenever $\mathbf{g}_i^\top \mathbf{g}_j < 0$. This ensures that task zz are adjusted to reduce destructive interactions. More recent approaches, such as CAGrad [78], further generalize conflict resolution by optimizing a common descent direction across tasks.

### Meta-Learning for Weight Optimization

Meta-learning frameworks offer an additional frontier for optimizing MTL. Guo et al. [79] proposed dynamic task prioritization, where validation performance guides the adjustment of task weights during training. These methods treat task weighting as a meta-optimization problem, enabling models to adapt dynamically to shifting task complexities and improving generalization.

### Evaluation Methodologies

Robust evaluation of MTL models requires metrics beyond average task accuracy. Authors in [80] framed MTL as a multi-objective optimization problem, highlighting the principle of Pareto optimality, wherein the performance of one task cannot be improved without causing a degradation in the performance of at least one other task. Evaluation strategies frequently involve trade-off analysis, per-task reporting, and visualization of task performance landscapes.

Overall, successful optimization in MTL hinges on dynamic, gradient-aware, and conflict-resolving strategies, combined with evaluation methodologies that account for task trade-offs and fairness.

## 2.6 Conclusion

This chapter provided a comprehensive overview of sentiment analysis, discussing major approaches including lexicon-based, machine learning-based, hybrid, and transfer learning methods, with special attention to the role of transformers. Challenges specific to sentiment analysis, particularly in low-resource settings, were highlighted. The discussion then focused on the Arabic language within the context of natural language processing (NLP), emphasizing its unique characteristics and significant differences from English. The chapter outlined the difficulties posed by Arabic language processing, examined linguistic features relevant to sentiment analysis, and presented examples of Arabic NLP applications. Morphological distinctions between Modern Standard Arabic and Arabic dialects were explored to demonstrate their impact on computational models. Arabic transformer models were introduced as a recent advancement to address these challenges effectively. Finally, the chapter explored Multi-Task Learning (MTL) for text classification, detailing its foundational principles and critical components, including shared and task-specific layers, encoder-decoder structures, parameter sharing strategies, and the use of multi-head attention within transformers. This chapter thereby established the theoretical and technical basis necessary for developing advanced, Arabic-specific sentiment analysis systems using deep learning and MTL frameworks.

# Chapter 3

# Literature review

## 3.1 Introduction

This chapter provides a comprehensive review of existing research in the field of sentiment analysis, with particular emphasis on applications to the Arabic language. Since most online content is unstructured, considerable effort is required to convert it into structured and meaningful information using a variety of analytical techniques and approaches. Recently, sentiment analysis has gained considerable attention within the research community as an effective method for knowledge representation and understanding public opinion. However, Arabic sentiment analysis presents unique challenges due to the language's morphological richness, dialectal variations, and complex syntactic structures. To address these difficulties, recent studies have proposed the integration of multi-task learning (MTL) frameworks, which enable simultaneous learning of related tasks to enhance model generalization and performance. This chapter critically examines the literature on sentiment analysis methodologies, with a specific focus on Arabic-language contexts, and discusses how MTL-based solutions are emerging as a promising approach for overcoming the linguistic and computational challenges associated with Arabic sentiment analysis.

## 3.2 Arabic Sentiment Analysis

Arabic sentiment analysis plays a crucial role in applications such as social media monitoring, market research, and political opinion mining. However, several linguistic challenges make sentiment classification in Arabic particularly complex. A major issue is the language's diglossia, where Modern Standard Arabic (MSA) coexists with numerous regional dialects that differ significantly in vocabulary, syntax, and morphology [81]. These dialects are often used in informal settings such as social media, creating variability that traditional models struggle to manage. Additionally, Arabic's morphological richness, where a single root can produce numerous word forms through complex derivational patterns, further complicates the sentiment analysis process [82].

Models that aim to address these issues must be capable of handling both MSA and dialectal Arabic while also managing orthographic variation and code-

switching phenomena frequently observed in online platforms. For example, social media texts often contain slang, phonetic spellings, and borrowed words, which conventional tokenization and embedding strategies may fail to capture effectively. Furthermore, the detection of sarcasm presents an additional layer of complexity in Arabic sentiment analysis, as sarcastic expressions can invert the intended polarity of a statement. This challenge has been highlighted in recent research emphasizing the need for sarcasm-aware sentiment classification models [83].

As the field matures, the development of dialect-aware and morphologically informed models is increasingly recognized as essential for accurate Arabic sentiment analysis. Addressing these challenges is crucial for building robust systems that can generalize across different genres, dialects, and linguistic phenomena in real-world Arabic text.

## 3.3 Lexicon-Based Approaches

Lexicon-based approaches have played a foundational role in Arabic sentiment analysis. A prominent example is the work of Badaro et al. [84], who introduced the ArSenL lexicon, a resource modeled after the English SentiWordNet. ArSenL integrates the Arabic WordNet (AWN) and the Standard Arabic Morphological Analyzer (SAMA), cross-referenced with the English SentiWordNet, and is available via a web-based graphical interface. Similarly, authors in [85] proposed an Arabic Sentiment Lexicon (ASL) developed from a seed list of positive and negative words. They employed a semi-supervised method to extend the lexicon to approximately 2,000 words—comprising 800 positive, 600 negative, and 600 neutral terms. However, since the resource was built using the general-purpose AWN, its coverage of domain-specific sentiment expressions remains limited and potentially insufficient for capturing the full range of sentiments present in Arabic-language reviews.

Abdul-Mageed et al. [86] advanced the field by constructing the SANA lexicon, a large-scale sentiment resource for Modern Standard Arabic (MSA) as well as Egyptian and Levantine dialects. SANA was developed through the integration of multiple approaches, including automatic translation from English sentiment lexica, manual curation, and statistical modeling. The final resource contained 224,564 entries. However, the presence of duplicate entries and inconsistent labeling reduced its overall effectiveness for downstream applications. A complementary approach was presented by Ayyoub et al. [87], who proposed an unsupervised sentiment classification system tailored to Arabic tweets. After collecting and preprocessing the data, they constructed a sentiment lexicon with numerical polarity scores ranging from 0 to 100. Sentiment classes were assigned as follows: scores above 60 were labeled positive, scores between 40 and 60 were considered neutral, and scores below 40 were classified as negative. The sentiment score of a sentence was calculated by aggregating the polarity scores of its constituent words. Despite achieving an accuracy of 86.89%, this method did not effectively handle dialectal variation, which limits its applicability across diverse Arabic linguistic contexts.

In response to the limitations of existing approaches in addressing dialectical nuances, authors in [88] proposed a four-phase framework. The first phase involved

selecting 300 sentiment seed words from the SentiStrength [89]lexicon. The second phase augmented the lexicon with synonyms of these seed words. In the third phase, a term frequency weighting method was employed to identify additional relevant terms. Finally, the fourth phase incorporated dialectal Arabic terms into the lexicon. Sentiment analysis was then performed using this enriched lexicon, relying on a basic polarity computation method that did not account for linguistic phenomena such as negation or intensification. The proposed system achieved an accuracy of 70.05% across various lexicon expansion stages but remained limited in effectively modeling dialectal intricacies.

Building on this line of research, authors in [90] introduced three distinct lexicon-based methods for Arabic sentiment analysis. One of these methods extended the basic lexicon model by incorporating mechanisms to handle contextual polarity specifically negation and intensification. By addressing these linguistic features, their enhanced model achieved a higher accuracy of 91.75%, significantly outperforming earlier methods such as that of [88].

Efforts to explore dialectal Arabic in sentiment lexicon development remain relatively sparse. Notable among these is the work of Mataoui et al. [91], who focused on Algerian dialectal Arabic, a variety characterized by frequent code-switching between Arabic and French. The researchers developed three lexicons: (1) a subset of an Egyptian dialect sentiment lexicon tailored to Algerian usage, (2) a manually compiled list of commonly used negative terms in Algerian Arabic, and (3) a list of intensifiers. Two system configurations were evaluated. The first operated at the phrase level and achieved satisfactory alignment with sentiment-labeled expressions. The second performed word-level analysis by applying normalization and tokenization processes, including language detection and stemming. Arabic tokens were stemmed directly, while non-Arabic tokens were first translated into Arabic before stemming. The system then matched stems with sentiment lexicon entries to compute semantic orientation. The authors manually annotated the sentiment polarity of 7,698 Facebook comments written in both MSA and Algerian dialect. Combining the two configurations, the model achieved an overall accuracy of 79.13%.

Further refining lexicon-based models, authors in [90] reaffirmed the benefit of accounting for contextual linguistic features. Their lexicon-based method integrating negation and intensification mechanisms confirmed earlier findings, yielding an accuracy of 91.75%. In contrast, Al-Moslmi et al. [92] highlighted that lexicon-based approaches are particularly advantageous for unlabeled data, as they facilitate automatic polarity labeling using pre-defined sentiment lexica. They demonstrated that sentiment could be effectively estimated by matching words and phrases in the input text to lexicon entries.

In another domain-specific application, authors in [93] developed a lexicon-based sentiment analysis model focused on Arabic tweets related to the Syrian civil war. Using a bag-of-words representation, tweets were classified as either positive or negative by comparing their content to an Arabic sentiment lexicon. Although the system achieved a classification accuracy of 68%, it lacked analysis of critical linguistic factors such as intensification and negation, and did not address dialectal

Arabic, limiting the comprehensiveness of its performance assessment.

## 3.4 Machine Learning-based Approaches

machine learning approaches have also played a foundational role in Arabic sentiment analysis. A prominent example is the work of Abdul-Mageed et al. [94] where they introduced the SAMAR system for subjective sentiment analysis, utilizing a diverse dataset that included Modern Standard Arabic (MSA) and various Arabic dialects words. Their work considered multiple domains, such as political, economic, sports, and entertainment news, drawn from sources like Wikipedia, tweets, online chats, and news forums. However, the SAMAR system showed limited effectiveness, particularly for tweets, where the sentiment classification yielded a low F-score of 49.41%.

Authors in [95] explored a four-level sentiment polarity classification using a dataset of approximately 815 Arabic comments collected from local Saudi Arabian online newspapers. Their model, trained with 620 comments and tested on 195, achieved an impressive accuracy of 85%. This work focused on handling negation and sentiment ambiguity in Arabic text, though it did not address the issue of irrelevant comment filtering during preprocessing.

In a related study, Itani et al. [96] focused on the creation of an annotated corpus from posts on Facebook pages like 'The Voice' and 'Al Arabiya.' Their goal was to improve natural language processing for Arabic dialect, specifically addressing sentiment classification. They employed multiple classifiers, such as decision trees (DT), support vector machines (SVM), and k-nearest neighbors (KNN).

Further advancements were made by Nabil et al.[97], who conducted a four-way sentiment classification of Arabic tweets, dividing them into objective, subjective negative, subjective positive, and mixed categories. Their dataset, comprising 10,006 manually annotated Arabic tweets, was processed using a variety of machine learning algorithms, including Naïve Bayes (NB), SVM, and stochastic gradient descent. They found that using n-grams as features for multi-way classification failed to yield satisfactory results, especially without preprocessing steps.

authors in [98] developed a sentiment analysis model for Saudi and Jordanian dialects. Their approach, which involved custom stop-word lists and light stemming, showed improvements by incorporating n-grams into the Bag-of-Words (BOW) representation. Their experiments indicated that the Maximum Entropy classifier performed best with trigrams.

Alomariet al. [99] analyzed 1,800 Jordanian tweets, categorizing them as negative or positive. By comparing Naïve Bayes and SVM classifiers, they found that SVM outperformed Naïve Bayes, achieving an F-score of 88.27%. Their research focused on different preprocessing techniques, including various stemming methods, and indicated that combining SVM with Term Frequency-Inverse Document Frequency (TF-IDF) worked most effectively.

Al-Rubaiee et al. [100] explored sentiment analysis for tweets from King Abdul-Aziz University students, using 2,000 tweets collected via the Twitter API. Their study applied light stemming, stop-word removal, and tokenization to Arabic text, but it faced challenges due to the relatively small dataset, suggesting that larger

datasets would improve the performance of their machine learning models.

Al-Horaibi et al. [101] proposed an emotion detection model for Arabic tweets, collecting 14,984 tweets using the Twitter API. They processed the data using Python libraries and applied classifiers such as Decision Tree (DT) and Naïve Bayes (NB). However, the use of English NLP tools led to suboptimal performance, as these tools were not well-suited for Arabic dialects.

Sghaier et al.[102] introduced a multi-algorithm sentiment analysis approach using KNN, SVM, and Naïve Bayes classifiers, achieving impressive accuracy rates (93.9% and 93.87% for SVM and NB, respectively). However, their study was limited by a small dataset of only 250 documents, and the lack of negation handling may have impacted results.

Baly et al.[103] delved into the challenges of sentiment analysis for Arabic tweets, focusing on dialectal variations and increased noise in the data. They created a typology of tweets to improve sentiment classification and employed SVM with POS tagging and lemmatization, yielding an accuracy of 55.70%.

Rahab et al. [104] worked on sentiment analysis for Algerian newspaper comments, experimenting with word-weighting strategies and classifiers such as KNN, SVM, and NB. The best results, with an accuracy of 75%, were achieved using Naïve Bayes combined with light stemming. Their study focused on a limited dataset, consisting of only 92 comments, which constrains the generalizability of the findings.

Mulki et al. [105] contributed to Arabic sentiment analysis during the SemEval International Workshop, tackling Twitter sentiment analysis as part of a subtask. Their study compared supervised and lexicon-based models and concluded that the supervised model provided the best performance, particularly when no stemming was used.

Maghfour et al. [106] focused on Moroccan Arabic and MSA Facebook comments. Their two-stage classification approach, incorporating light stemming for dialectical texts, improved classification accuracy and minimized stemming errors. However, they noted that this method would face difficulties in larger and more diverse multi-dialect datasets.

Sayed et al. [107] developed a multidimensional sentiment analysis system for Arabic using a dataset of 6,318 reviews from Booking.com. Their experiments with nine different classifiers, including KNN, RC and SVM, revealed that Ridge Classifier (RC) performed the best in terms of recall, precision, and F1 score. They also highlighted the importance of preprocessing steps, such as stemming and stop-word removal, in enhancing classification accuracy.

Finally, Baly et al. [108] represented ArSentD-LEV, a multi-topic corpus specifically designed for target-based sentiment analysis in Arabic Levantine tweets. The dataset addresses the linguistic complexity inherent in Levantine dialects and offers fine-grained annotations that go beyond document-level sentiment. It captures sentiment directed at specific entities or targets within tweets, facilitating more nuanced analysis. ArSentD-LEV overcomes challenges related to dialectal variation, informal language, and limited resources by focusing on a specific regional dialect, thereby enabling more accurate sentiment detection and opinion mining in a real-world social media context.

## 3.5 Deep Learning-based Approaches

Deep learning has demonstrated transformative capabilities across a range of domains, including Artificial Intelligence, Computer Vision, and the Internet of Things [109]. In the field of Natural Language Processing (NLP), the impact of deep learning methods has been particularly significant, especially in tasks such as sentiment analysis. With Arabic being a morphologically rich and dialectally diverse language, the adoption of deep learning methods for sentiment classification has provided promising outcomes compared to traditional machine learning techniques.Two foundational components in deep learning-based sentiment analysis are the representation of textual data and the modeling of its linguistic features. Text is commonly represented using traditional approaches like Bag-of-Words (BoW), or more advanced word embedding techniques such as Word2Vec [110] and GloVe [111], which are designed to capture both semantic and syntactic nuances. These embedding methods are particularly critical for effectively processing the intricate linguistic structures and variability found in Arabic dialects. In Arabic sentiment analysis, deep learning models are typically categorized into three major architectures: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid neural models. Each of these approaches brings distinct advantages and limitations, particularly in addressing the challenges posed by the linguistic features of Arabic.

### CNN-based Models

Convolutional Neural Networks (CNNs) have been successfully adapted for sentiment analysis task. Kim's seminal work [112] demonstrated that CNNs, when combined with pre-trained word embeddings, could perform competitively on various sentence classification tasks. In the Arabic context, CNNs have been used effectively to capture local patterns in text, such as key phrases or expressions indicative of sentiment, while maintaining computational efficiency. For instance, Alayba et al. [113] applied CNNs to Arabic tweets and demonstrated their capability to outperform traditional classifiers in binary and ternary sentiment classification tasks. Further advancements include deeper CNN architectures, such as the Very Deep Convolutional Neural Network (VDCNN) proposed by Conneau et al. [114], which has inspired deeper modeling in Arabic texts. In a similar vein, one-dimensional CNNs have also been applied to Arabic sentiment datasets by Johnson and Zhang [115], revealing that convolutional layers can effectively model hierarchical features in Arabic scripts. Nevertheless, CNNs often struggle to capture long-range dependencies in sentences, a notable limitation when dealing with Arabic's free word order and syntactic variability.

### RNN-based Models

Recurrent Neural Networks (RNNs), and particularly their variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), are widely recognized for their ability to model sequential data [116, 117]. This capability

makes them especially suitable for processing natural language where context and word order significantly affect meaning. In Arabic sentiment analysis, RNNs have been instrumental in handling longer texts, such as customer reviews or opinion articles, where sentiments are context-dependent and influenced by surrounding words. Tang et al. [118] demonstrated that hierarchical RNN architectures, incorporating both sentence-level and document-level features, improve sentiment classification performance. This approach has been adapted in Arabic by various researchers to capture the intricacies of Modern Standard Arabic (MSA) and regional dialects. For instance, Alharbi et al. [119] utilized bidirectional LSTMs on Arabic tweets, enabling the model to consider both past and future contexts, which proved beneficial in detecting subtle sentiment cues. However, despite their advantages, RNNs face limitations such as difficulty in modeling long sequences and high computational costs. These shortcomings have led researchers to explore alternative or complementary architectures, including attention mechanisms and Transformer-based models, to better handle the structural and semantic variability in Arabic.

### Hybrid Neural Network Models

Hybrid models combine multiple neural architectures or integrate other mechanisms such as attention layers, ensemble techniques, or meta-learning frameworks to enhance the performance of sentiment analysis systems. In Arabic sentiment analysis, hybrid approaches have been particularly valuable due to the language's diglossia and morphological richness, which often require more sophisticated modeling than what single-architecture models can offer. For example, Yang et al. [120] introduced a hierarchical attention network, which has been used in Arabic sentiment research to prioritize sentiment-bearing words and phrases within larger textual contexts. The attention mechanism, by explicitly highlighting influential terms, addresses the RNN's limitation in treating all context words equally.

### Challenges and Future Directions

While deep learning approaches offer considerable promise for Arabic sentiment analysis, several key challenges persist. First, most deep learning models require large annotated corpora, yet publicly available Arabic sentiment datasets especially for dialects remain limited. Manual annotation is costly and time consuming, and automatic labeling often introduces noise due to the lack of reliable sentiment lexicons for Arabic dialects. Second, deep learning models often require high computational resources, limiting their accessibility for researchers without access to GPUs or specialized hardware. Third, the generalization of models across different dialects and domains remains problematic. A model trained on Levantine Arabic tweets, for instance, may not perform well on Gulf Arabic reviews due to lexical and syntactic differences. In conclusion, deep learning has significantly advanced Arabic sentiment analysis, enabling more accurate and context-aware models. However, future research must continue to address challenges related to data scarcity, computational demands, and cross-dialect generalization. The integration of hybrid models, attention mechanisms, and transfer learning techniques will be critical

in overcoming these limitations and advancing the state of Arabic sentiment analysis.

## 3.6 Transfer Learning-based Approaches

Transfer learning has revolutionized the field of natural language processing (NLP), providing robust solutions for tasks in low-resource languages such as Arabic. The core principle of transfer learning is the adaptation of knowledge learned from a resource rich source domain to a related target domain with limited labeled data [121]. This framework has proven particularly beneficial for Arabic sentiment analysis, a field that faces significant challenges due to the linguistic complexity of Arabic, its rich morphology, and its diglossic nature. Arabic comprises Modern Standard Arabic (MSA) and a multitude of dialects that differ significantly at the phonological, lexical, and syntactic levels [122]. These characteristics complicate supervised learning, as manually annotated data is scarce, especially for dialects. Transfer learning has thus become a key methodology to bridge this resource gap and enhance performance in Arabic sentiment analysis.

Historically, early applications of transfer learning in sentiment analysis were grounded in instance-based transfer, which involves reusing or reweighting specific source domain examples based on their relevance to the target domain [123]. These methods aim to align data distributions through importance sampling or instance selection. For Arabic, authors in [124] implemented instance reweighting mechanisms to adapt MSA-trained models to dialectal datasets. While these approaches showed moderate success, they often struggle with scalability and robustness when applied to neural architectures. Furthermore, they rely on accurate estimation of instance similarity, which is particularly difficult in morphologically rich languages such as Arabic.

A more scalable strategy in deep learning settings is feature-representation transfer, which involves learning abstract features that generalize well across domains. With the emergence of word embeddings like Word2Vec [125], GloVe [126], and fastText [127], cross-domain feature sharing became feasible. Arabic NLP benefited from cross-lingual embedding techniques such as MUSE [128], which align multilingual vector spaces and allow for leveraging resources from high-resource languages to support Arabic tasks. Recent transformer-based models further extend this principle, AraBERT [57] represents one of the earliest BERT-based models specifically adapted for the Arabic language. It was pretrained on a corpus exceeding 200 million sentences sourced from a wide range of Arabic texts. The model demonstrated state-of-the-art performance across multiple sentiment analysis and text classification benchmarks. AraBERT effectively illustrates the strength of feature representation transfer, as it is capable of capturing intricate syntactic and semantic patterns that are shared across various Arabic dialects.

The most dominant approach today is parameter-based transfer learning, where large pretrained language models (PLMs) are fine-tuned on downstream tasks. BERT [29], GPT [129], and their multilingual variants, such as mBERT and XLM-R [130], embody this approach. In Arabic sentiment analysis, both multilingual and monolingual PLMs have been explored. MARBERT [26], for

example, was trained on 1 billion Arabic tweets, including a large representation of dialectal Arabic. It outperformed earlier models on sentiment and dialect identification tasks, demonstrating the utility of domain-specific parameter transfer. Likewise, QARiB [131] and AraELECTRA [132] have been introduced as Arabic-specific transformer models with enhanced performance on sentiment and emotion classification benchmarks.

Cross-lingual parameter transfer has also been investigated in contexts where large-scale Arabic training data is unavailable. In this approach, models pretrained on high-resource languages are adapted to Arabic through either zero-shot or few-shot transfer learning. XLM-R [130], trained on 100 languages, has demonstrated competitive performance in Arabic sentiment classification by leveraging shared subword representations across languages [133]. Authors in [134] found that multilingual PLMs such as XLM-R can outperform Arabic-specific models when trained with proper tokenization and dialect-specific preprocessing. These results suggest that under certain conditions, cross-lingual transfer can offer competitive or even superior performance, especially when combined with domain-specific fine-tuning.

An emerging and promising transfer approach is multi-task learning (MTL) [64], in which multiple related tasks are trained jointly, encouraging the model to learn generalized representations. In Arabic sentiment analysis, MTL has been used to integrate sentiment classification with sarcasm detection, dialect identification, or emotion recognition. For instance, Abdelrahman and mona [135] introduced SAIDS, a system that jointly predicts sentiment, sarcasm, and dialect in Arabic tweets. By leveraging dialect and sarcasm predictions as auxiliary tasks, the model enhances sentiment analysis performance. Similarly, studies by Brahim et al.[136] showed that training models to perform both sarcasm detection and sentiment analysis led to more robust performance on informal Arabic text.

Domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT) represent more refined approaches to transfer learning. These techniques involve continuing the pretraining phase on unlabeled data that is either domain-specific or task-specific, thereby narrowing the domain discrepancy between the source and target data distributions [137]. For example, continuing AraBERT's pretraining on Arabic tweets (TAPT) prior to sentiment analysis fine-tuning has shown significant gains in classification accuracy [57]. Similarly, [26] applied (DAPT) to MARBERT using large volumes of dialectal tweets to better adapt to user-generated social media content. These refinements enable models to specialize their general knowledge in ways that are directly relevant to the target application.

Adversarial transfer learning [138] is also gaining traction in Arabic NLP as advanced strategies for learning domain-invariant and task-discriminative representations. Adversarial learning employs a domain classifier that penalizes the model for encoding domain-specific features, thereby promoting generalizable representations [139]. This approach has been adopted in Arabic sentiment tasks to mitigate distributional mismatches between MSA and dialects [140].

Despite the progress, several challenges persist. First, dialectal diversity remains a major obstacle. Arabic dialects vary significantly across regions, and their digital

presence (e.g., tweets, comments) is often noisy and unstandardized [141]. Many pretrained models, even those tailored to Arabic, perform unevenly across dialects. MARBERT attempts to address this through dialectally diverse pretraining data, but dialect-specific fine-tuning remains necessary. Second, code-switching, which is common in North African and Levantine regions, poses challenges to monolingual models, and cross-lingual transformers like XLM-R that have only partially mitigated this issue [142]. Third, evaluation benchmarks are not standardized across Arabic variants, making it difficult to fairly compare models. Datasets such as ASTD [97], ArSenTD-Lev [108], and LABR [143] differ widely in dialect, domain, and annotation schemes.

Looking ahead, researchers are increasingly combining transfer learning with semi-supervised and unsupervised methods to overcome data bottlenecks. Weak supervision, pseudo-labeling, and data augmentation techniques such as back-translation are being used to generate synthetic labeled data for fine-tuning [144]. For example, multimodal transfer learning that integrates textual, visual, and acoustic signals is gaining attention for sentiment tasks on social media, particularly platforms like TikTok or Instagram where emojis, images, and slang interact with textual sentiment [145].

Transfer learning has become indispensable in Arabic sentiment analysis. From early instance reweighting techniques to sophisticated PLMs like BERT, XLM-R and GPT the ability to leverage external knowledge has significantly improved the performance and generalizability of sentiment classification systems. While challenges such as dialectal variation, data sparsity, and noise persist, new paradigms such as multi-task learning, DAPT offer promising pathways forward. Continued innovation in transfer learning will undoubtedly play a central role in developing scalable, accurate, and dialect-aware sentiment analysis systems for Arabic in the years to come.

## 3.7 Multi-Task Learning-based Approaches

Multi-Task Learning (MTL) [64] has become a prominent technique in natural language processing (NLP), particularly in resource constrained or linguistically complex environments. In the case of Arabic Sentiment Analysis (ASA), MTL has proven to be a viable approach to mitigating challenges posed by dialectal diversity, ambiguity in sentiment-laden expressions, and the difficulty of detecting sarcasm. Unlike traditional single-task learning (STL) approaches, MTL allows a model to jointly learn related tasks such as sentiment classification and sarcasm detection by sharing representations across them. This strategy can lead to better generalization, faster convergence, and robustness against overfitting, especially when annotated data is limited, as is often the case with Arabic dialect corpora.

The foundational principles of MTL were introduced by Caruana [59], who argued that tasks sharing common structures could benefit from joint optimization by leveraging shared inductive biases. When models are trained concurrently on multiple tasks, the shared parameters encourage the learning of more general and transferable features. This is particularly useful in ASA, where sentiment cues

may depend heavily on dialectal context and may be confounded by ironic or sarcastic tone. Dialectal Arabic introduces lexical and syntactic variation that challenges sentiment classifiers trained on Modern Standard Arabic (MSA), often resulting in reduced accuracy [146]. Therefore, models that can simultaneously recognize dialect and infer sentiment are more effective than models trained on sentiment labels alone. Moreover, the presence of sarcasm within a text further complicates sentiment analysis, as conventional models typically fail to detect sarcastic expressions, thereby reducing classification accuracy.

In recent years, transformer-based language models such as BERT [29], RoBERTa [147], and XLM-R [130] have been adapted to Arabic, giving rise to Arabic-specific variants like AraBERT [57], MARBERT [26], and CAMeLBERT [148]. These models provide deep contextualized embeddings that are particularly suited for downstream tasks, including ASA. While most of these models were originally applied in single-task settings, several studies have recently explored their utility in multi-task configurations. For instance, authors in [136] organized a shared task on Arabic sarcasm and sentiment detection, demonstrating that sarcasm detection is not only a challenging task on its own but also beneficial when integrated with sentiment analysis in a multi-task learning framework. Models submitted to this task showed that joint modeling improved sentiment prediction, especially in cases where sarcasm altered the polarity of the sentence.

Building on these insights, the authors' findings in [135] confirm that auxiliary tasks provide complementary information that enhances the model's ability to disambiguate sentiment in complex expressions. By learning dialect-specific nuances and recognizing sarcastic cues, the model achieved higher F1 scores on benchmark datasets compared to its single-task counterparts. This underscores the utility of MTL in contexts where sentiment labels alone do not provide sufficient supervision due to variability in language usage across different dialects.

Despite these promising results, challenges remain. One persistent issue is the scarcity of large-scale, high-quality annotated datasets that contain aligned labels for multiple tasks. While datasets like ArsentD-Lev [108], ArSarcasm [149] and ASTD [97] have supported research in sentiment and sarcasm, few corpora offer multi-task annotations that cover dialect, sentiment, and other subtasks simultaneously. Consequently, some researchers rely on distant supervision, weak labeling, or synthetic data augmentation, which introduces noise and may limit generalization.

Multi-Task Learning has demonstrated clear benefits in Arabic Sentiment Analysis by enabling models to learn more comprehensive and context-sensitive representations. As research progresses, the development of unified multi-task datasets and further exploration of task selection strategies will be essential to unlock the full potential of MTL in Arabic NLP.

## 3.8 Conclusion

This chapter has presented a comprehensive review of the literature pertinent to Arabic sentiment analysis, tracing the evolution of methodological approaches and the progression of technological frameworks in the field. The chapter commenced

with an overview of Arabic sentiment analysis, emphasizing the linguistic challenges posed by morphological complexity, and dialectal variation, all of which necessitate specialized tools and techniques.

Subsequently, a focused review of lexicon-based approaches was provided, detailing early efforts that relied on curated sentiment dictionaries and rule-based systems. While foundational, these methods were shown to be limited in scalability and adaptability, particularly in processing dialect-rich and context-dependent Arabic data. This led to an examination of traditional machine learning techniques, where classifiers such as SVMs and Naïve Bayes were applied to sentiment-labeled corpora. These methods demonstrated modest performance improvements but remained heavily reliant on feature engineering and domain-specific preprocessing.

The chapter then transitioned to deep learning approaches, highlighting the advent of neural network architectures such as CNNs, RNNs, and LSTMs. These models enabled automatic feature extraction and demonstrated significant gains in sentiment classification accuracy. Nonetheless, their performance was constrained by limited annotated Arabic corpora and the challenge of generalizing across dialects.

The emergence of transfer learning represented a paradigm shift, allowing pre-trained transformer models such as AraBERT, MARBERT, and multilingual BERT to be fine-tuned for Arabic sentiment tasks. This section underscored the effectiveness of these models in leveraging large-scale language representations, even in low-resource settings, and discussed key advancements in cross-lingual and cross-dialectal transfer.

Finally, the review addressed multi-task learning (MTL) frameworks, which jointly model sentiment analysis alongside auxiliary tasks. These architectures were found to enhance performance by capturing shared linguistic representations and mitigating task-specific overfitting. Notably, MTL has shown promise in addressing the nuances of Arabic sentiment analysis, particularly in informal and user-generated content.

# Chapter 4

# Methodology / Proposed Methods

## 4.1 Introduction

In recent years, Arabic sentiment analysis has emerged as a critical area of research within natural language processing (NLP), driven by the increasing volume of user-generated content across social media platforms. Arabic sentiment analysis is a challenging task due to the language's complex morphology, rich lexical variation, and the lack of standardized orthography across dialects. These linguistic intricacies are further complicated by the presence of sarcasm, informal expressions, and dialectal ambiguity, which significantly degrade the performance of conventional text-based classification models. To address these challenges, this chapter proposes two new architectures that aim to enhance the robustness and accuracy of Arabic sentiment classification. The first approach leverages multimodal learning to integrate multimodal data sources textual, categorical, and numerical within a unified framework. By utilizing the ArsenTD-Lev dataset, which includes some specific annotations, this approach exploits the potential of multimodal representations to mitigate information loss inherent in text-only models. The second approach explores the use of Multi-Task Learning (MTL) mechanism for joint optimization across three tasks: Arabic sentiment classification (ASA), Arabic sarcasm detection (ASD), and Arabic dialect identification. By sharing a common transformer encoder and allowing for task-specific decoders, the model captures shared linguistic features while maintaining task-level distinctions. This joint learning strategy is particularly effective in low-resource settings and enhances the model's ability to disambiguate sentiment in sarcastic or dialectally diverse contexts. These two proposed architectures aim to push the boundaries of current methodologies in Arabic sentiment analysis by introducing strategies that are both scalable and adaptable to real-world data complexity.

In this chapter, Section 4.2 provides an in-depth overview of Enhancing Arabic Sentiment Analysis through Multimodal-Toolkit approach , while Section 4.3 focuses on the architecture of the Multi Task Learning for Multi-dialect Arabic Sentiment Classification and Sarcasm Detection approach.

## 4.2 Enhancing Arabic Sentiment Analysis through Multimodal-Toolkit: A Study on Levantine Arabic Dataset

In this work, we aimed to enhance the accuracy of Arabic sentiment classification by incorporating a multimodal learning approach. We utilized a toolkit designed to integrate multimodal data including text, categorical, and numerical features for both classification and regression tasks. As shown in figure 4.1 our framework is primarily based on Transformer models, with an additional combining module in Table 4.2 that fuses the outputs of the Transformer encoder with external features. This fusion generates rich, multimodal representations that are subsequently passed to downstream classification or regression layers.

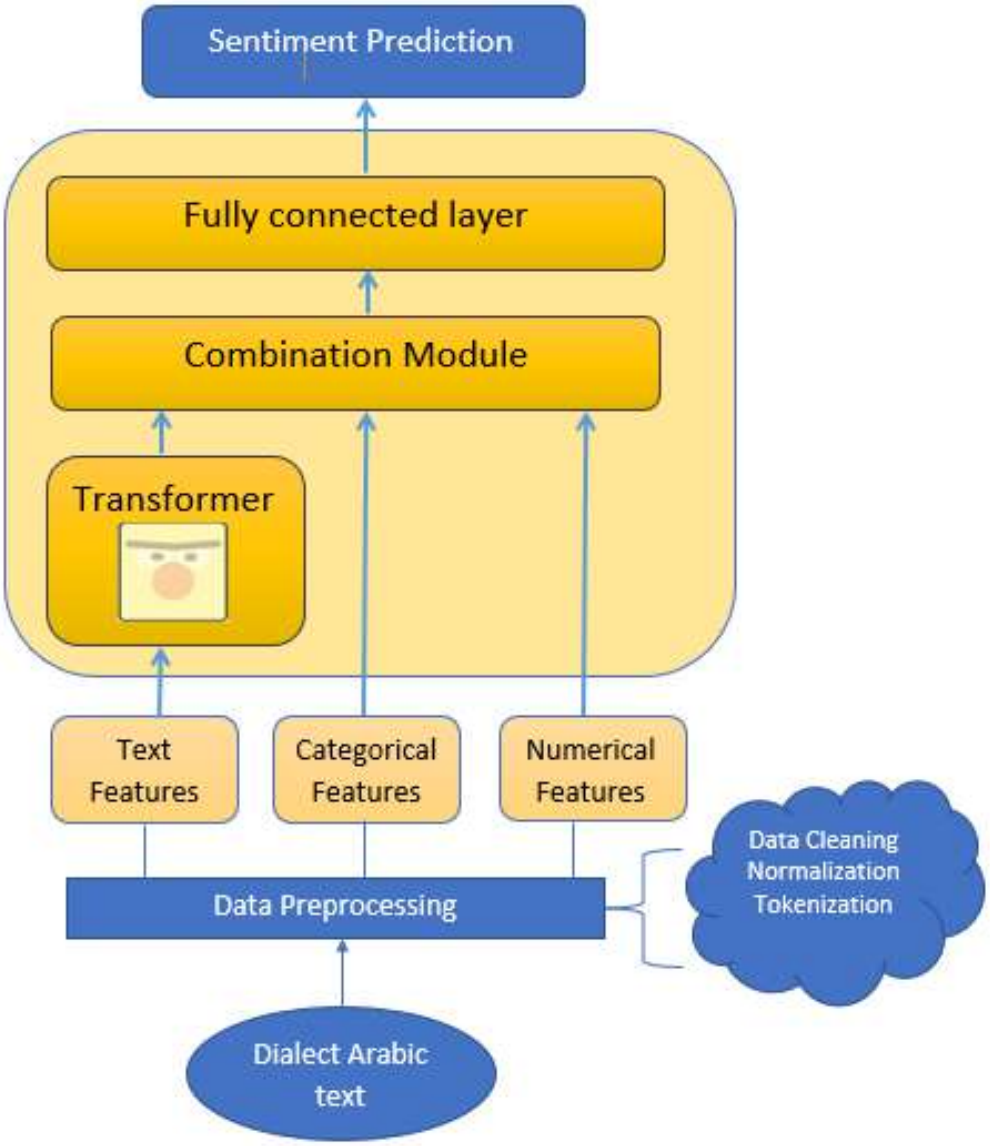



**Figure 4.1:** The proposed architecture using the Multi-modal toolkit.

### 4.2.1 Data Preparation

In this work, we focused on dialectal Arabic, specifically the ArsenTD-Lev dataset. As shown in table 4.1 the dataset includes not only textual content but also structured information such as categorical and numerical features. This multimodal nature enhances the suitability of the dataset for comprehensive sentiment analysis, particularly for Levantine dialects. These dialects often exhibit significant deviations from Modern Standard Arabic in both form and expression, making them challenging for traditional models. By incorporating both textual and structured data, the dataset provides a richer, more nuanced understanding of sentiment, especially in the context of dialectal variations.

| Tweet | Country | Topic | Sentiment | Expression | Target |
|---|---|---|---|---|---|
| أنا أؤمن بأن الانسان ينطفئ جماله عند ابتعاد من يحب حتى بريق العيون يختفي… | Lebanon | Personal | Negative | 2 | بريق العيون |
| إن لله وإن إليه راجعون اللهم أجرنا في مصيبتنا وأخلفنا… وأهلاً خيرًا من اهلها | Jordan | Personal | Negative | 1 | اللهم أجرنا في مصيبتنا |
| لا نخلو من ضغوطات الحياة… فنحن نعيش على أرض أعدت للبلاء ولم يسلم منها حتى الأنبياء… توكل على الله دائماً وكن مطمئناً وواثقاً بالله | Palestine | Personal | Neutral | 0 | None |
| وأن أبتدأ مسائي فيك دريت أنه مساء العافية. مساء الخير | Palestine | Personal | Very_Positive | 2 | دريت أنه مساء العافية |
| اللهم إنا نسألك في يوم الجمعة خير يشبه ألمطر وفرحة تمحي كل حزن ، وفرج لكل صابر ، … إنك على كل شيء قدير | Palestine | Religious | Very_Positive | 1 | يوم الجمعة |

**Table 4.1:** An example of a ArsentD-Lev classification dataset. Each row is a data point consisting of text, categorical features, and numerical features [108].

### 4.2.2 Model Architecture

The proposed architecture as shown in figure 4.1 incorporates a BERT-based encoder to extract high-level features from textual input. In addition to textual data, categorical and numerical features are preprocessed and integrated using a combining module 4.2. This module takes the encoded text features (x), categorical features (c), and numerical features (n) to generate a unified multimodal representation (Z). The combined representation is then passed through a fully connected layers to produce the final sentiment prediction. This architecture ensures the effective incorporation of multimodal data sources by allowing the model to leverage the complementary strengths of textual, categorical, and numerical inputs for improved sentiment classification performance.

**Table 4.2:** The included combining methods in the combining module. Uppercase bold letters represent 2D matrices, lowercase bold letters represent 1D vectors. $b$ is a scalar bias, WM represents a weight matrix, and || is the concatenation operator.

| ***Combine Feature Method*** | ***Equation*** |
|---|---|
| Text only | $Z = x$ |
| Concat | $Z = x\|\|c\|\|n$ |
| Gating on categorical and numerical features and then sum (Gating) [150] | $Z = x + \alpha H$<br>$H = g_c \odot (WM_c c) + g_n \odot (WM_n n)$<br>$+b_H$<br>$\alpha = min(\frac{\|\|x\|\|_2}{\|\|h\|\|_2}\beta, 1)$<br>$g_i = R(WMg_i[i\|\|x + b_i)$<br>where $\beta$ is a hyperparameter<br>and R is a non-linear activation function |
| Weighted feature sum on text, categorical, and numerical features (Weighted Sum) | $Z = x + wm_c.WM_c c + wm_n.WM_n n$ |

## 4.3 Multi Task Learning for Multi-dialect Arabic Sentiment Classification and Sarcasm Detection

In this study, we propose a model based on a Multi-Task Learning (MTL) framework designed to concurrently address three essential Arabic NLP tasks: sentiment classification, sarcasm detection, and dialect identification. The goal is to leverage shared representations to enhance generalization and performance across these related tasks. Among the various MTL strategies discussed in the literature, the hard parameter sharing which has been adopted in this work is the most prevalent approach in deep learning and NLP. The hard parameter sharing parameter allows for the joint learning of shared parameters across tasks, thereby facilitating efficient training and promoting the extraction of generalized linguistic features. To assess the effectiveness of this approach, we conduct a comparative evaluation between MTL-based models and their Single-Task Learning (STL) counterparts. The architecture of the proposed MTL network is depicted in Figure 4.3.

Sentiment classification and sarcasm detection are closely related tasks, as sarcasm often modifies the intended sentiment of a given expression [151]. To

capitalize on this relationship we propose an MTL framework designed for joint training on Arabic Sentiment Analysis (ASA), Arabic Dialect Identification, and Arabic Sarcasm Detection (ASD). This framework is also designed to address challenges arising from lexical ambiguity, where identical words may convey different meanings across dialects. The overall framework for multi-dialect Arabic sentiment and sarcasm prediction is presented in Figure 4.2.

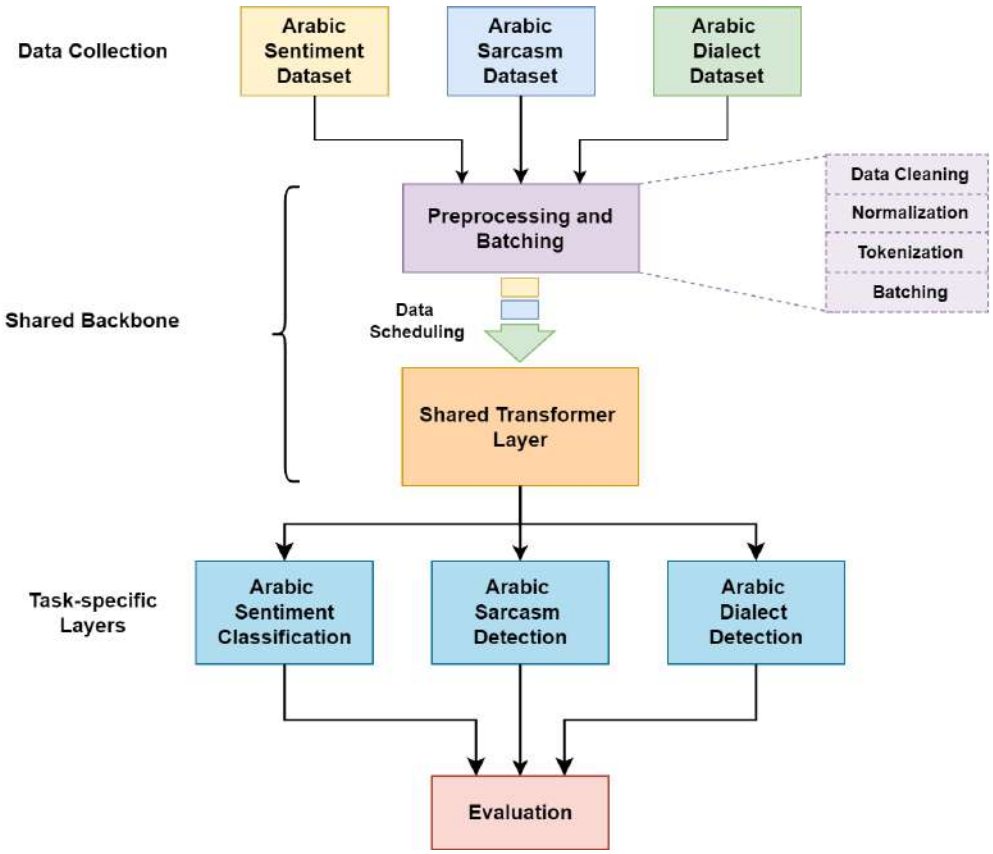



**Figure 4.2:** The structure of the suggested MTL framework.

### 4.3.1 Proposed model

In our proposed framework, a transformer encoder serves as a shared representation layer to extract contextualized embeddings from the input across all tasks 4.3. The input sequences are tokenized using the WordPiece tokenizer [152], which segments the text into subword units. Special tokens are appended to each sequence: the classification token [CLS] is placed at the beginning of the sequence, and the separator token [SEP] is appended at the end of the sequence to indicate the boundary of the input. The [CLS] token embedding is used for classification purposes, while [SEP] denotes the conclusion of the input sequence. Let $Y_t$ represent the count of unique labels, and $C_t$ represent the count of unique classes. Given a sentence $S = [CLS], s_1, s_2, \ldots, s_n, [SEP]$ as a tokenized input, the transformer encoder module will generate an output $X = E_C LS, E_2, E_3, \ldots, E_N$ of size $[D_o, N] \epsilon R^{D_o x N}$, where $D_o = 768$ and $N = 2 + n$ and $D_o$ denotes the final hidden layer dimension

of the transformer encoder.

$$X = TransformerEncoder(S) \tag{4.1}$$

Figure 4.3 presents the architecture of our model, which employs a transformer encoder as a shared representation layer across three classification tasks: Arabic sentiment analysis, Arabic sarcasm detection, and Arabic dialect identification. This shared encoder is subsequently followed by task-specific fully connected layers, each tailored to the individual objectives of the respective tasks.

The model receives as input a sentence $S$, which is first processed using the WordPiece tokenizer. This tokenizer decomposes the text into subword units to effectively manage morphological richness and out-of-vocabulary terms. These subword units are then mapped to token embeddings of dimensionality 1024 (assuming AraBERT model). The resulting embeddings are fed into the shared transformer layer.

The transformer layer consists of 16 attention heads and 24 transformer layers, each with 1024 hidden units per layer. This shared layer is designed to encode the contextual information from the input sentences. It produces a sequence of hidden states, denoted as $X$, where each token in the input sequence is represented by a contextual embedding of size 1,024.

For each classification task, the output $X$ from the shared transformer encoder is passed through a task-specific fully connected layer (FCL), also referred to as a dense layer. These layers project the high-dimensional contextual representation $X$ into a lower-dimensional, task-specific sentence-level embedding, denoted as $Z_i$ for task $i$. This transformation is formally represented in Equation 4.2.

$$Z_i = FCLayer_i(X) \tag{4.2}$$

An activation function is subsequently applied to the sentence-level representation $Z_i$ to produce the final task-specific classification output. In this framework, the softmax activation function is employed for the Arabic sentiment classification and Arabic dialect identification tasks, enabling multi-class predictions. For the Arabic sarcasm detection task, a sigmoid activation function is utilized to support binary classification. These operations are formally defined in Equations 4.3, 4.4, and 4.5.

$$S_{Sentiment} = Softmax(Z_1) \tag{4.3}$$

$$S_{Sarcasm} = Sigmoid(Z_2) \tag{4.4}$$

$$S_{Dialect} = Softmax(Z_3) \tag{4.5}$$

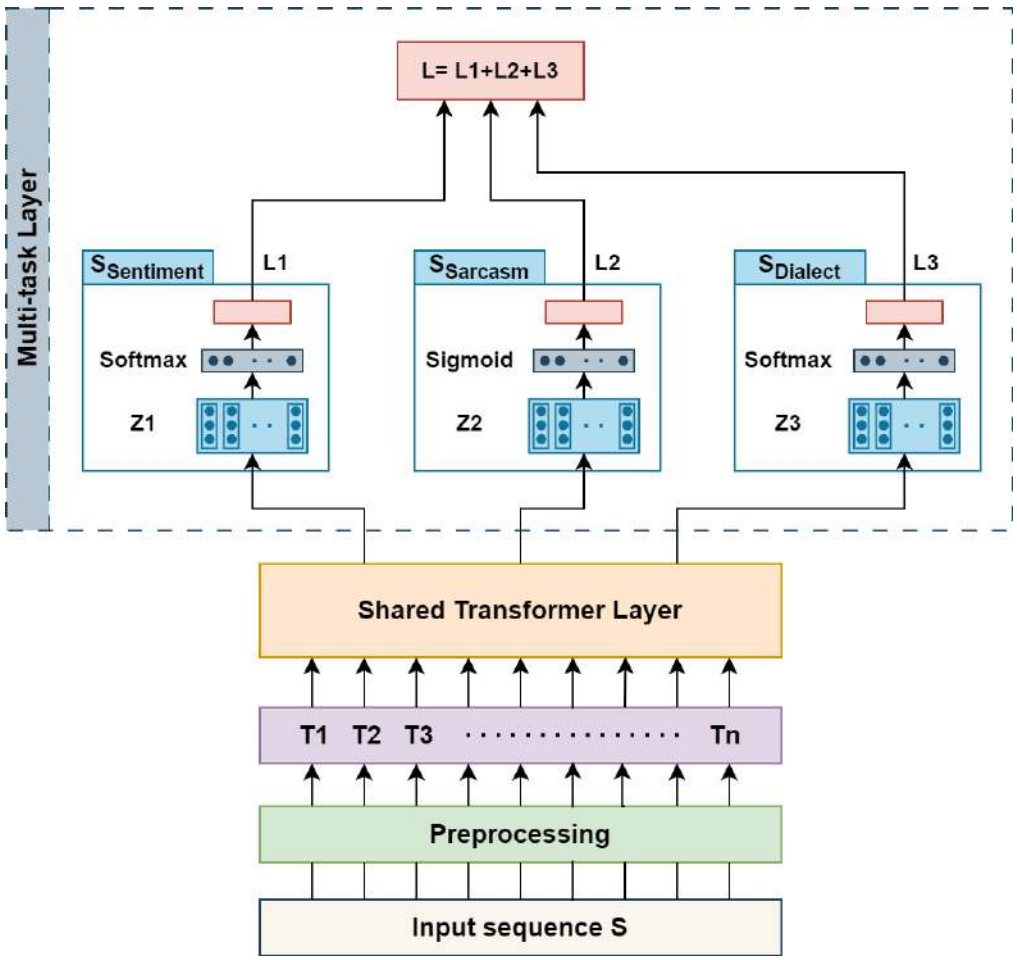



**Figure 4.3:** The Structure of our Multi-Task Learning network.

### 4.3.2 Loss Function

Our framework setup enables the second and third tasks to contribute information for the primary task training by calculating the model's loss using Equation 4.6.

$$L = \sum_{(x,y)\in\Omega_1} L_1(x,y) + \sum_{(x,y)\in\Omega_2} L_2(x,y) + \sum_{(x,y)\in\Omega_3} L_3(x,y) \tag{4.6}$$

Where $L_1$ is the loss for the primary task, $L_2$ and $L_3$ identified as the loss for the secondary and the third tasks, respectively. $L$ is used to calculate the total loss for every single sentence in the dataset $\Omega_i$.

We employed the binary cross-entropy loss for Arabic sarcasm detection, as presented in Eq. 4.7.

$$H(y,p) = -[(1-y).\log(1-p) + y.\log(p)] \tag{4.7}$$

Where $p$ is the predicted probability that the input belongs to class 1, and $y$ is the true label (0 or 1). We used another cross-entropy loss formula in the case of Arabic sentiment classification and Arabic dialect identification with C classes, as shown in Eq. 4.8:

$$H(y,p) = -\sum_{i=1}^{C} y_i . \log(p_i) \tag{4.8}$$

Where y is a one-hot encoded vector representing the true class (e.g., [0,1,0] for the second class in a 3-class problem), p is a vector of predicted class probabilities for each class and $log(p_i)$ computes the natural logarithm of the predicted probability for class i.

The utilization of the Adam optimizer is employed within our framework to enhance the performance of the model. In each epoch, the suggested algorithm computes the $L_i$ gradient for each batch to adjust the parameters.

## 4.4 Conclusion

In this chapter, we presented two new architectures aimed at addressing key challenges in Arabic sentiment analysis: (1) Enhancing Arabic Sentiment Analysis through Multimodal-Toolkit, and (2) Multi-Task Learning for Multi-Dialect Arabic Sentiment Classification and Sarcasm Detection. The first architecture integrates the Multimodal Toolkit to process heterogeneous data sources, with a particular focus on dialectal Arabic datasets. Specifically, we employed the ArsenTD-Lev dataset, which encompasses not only textual data but also categorical and numerical features, making it particularly well-suited for our study. By incorporating multimodal data, our aim was to enhance the performance of sentiment analysis models beyond the capabilities of text-only approaches. In the second architecture, we employed a Multi-Task Learning (MTL) algorithm built upon a pre-trained Arabic language model to jointly perform sentiment classification and sarcasm detection. We claim that having the ability to accurately detect sarcasm is essential for improving the overall reliability of sentiment classification, particularly in dialectal and informal Arabic contexts.

Although detailed experimental evaluations and comparisons with state-of-the-art methods will be presented in the following chapter, the proposed architectures establish a solid foundation for advancing Arabic sentiment analysis within the field of natural language processing. Future research will aim to further refine both models, enhancing their scalability and extending their applicability to a broader spectrum of Arabic dialects and diverse real-world scenarios. Such advancements are anticipated to improve the generalizability and robustness of the models, thereby increasing their effectiveness in sentiment classification tasks.

# Chapter 5

# Experimental analysis

## 5.1 Introduction

Arabic sentiment analysis remains a challenging task due to the language's complexity. To address its limitations, this chapter presents two complementary studies aimed at enhancing dialectal Arabic sentiment analysis by exploring both multimodal data integration and multi-task learning strategies.

The first study investigates the effect of integrating tabular (structured) features with textual data to improve sentiment classification performance in dialectal Arabic. Specifically, we utilize the ArSenTD-Lev dataset to evaluate whether integrating tabular features with text-based embeddings can further enhance classification outcomes. Through a series of controlled experiments, we compare the results of combining approaches with the state-of-the-art text-only baseline models' results. The results provide empirical evidence that structured data can complement deep contextual embeddings by supplying syntactic and statistical cues that are often underrepresented in purely textual inputs.

The second study focuses on leveraging multi-task learning (MTL) using transformer-based architectures to jointly model sentiment classification, sarcasm detection, and dialect identification. Recognizing the interdependence of these tasks in real-world language understanding, where sarcasm often alters sentiment, and dialect affects both. In this study, we proposed and evaluated several MTL configurations on ArSarcasm, ArSentD-Lev, NADI and ASTD datasets. Our results demonstrate that shared learning across related tasks allows models to extract richer, more generalizable representations, leading to improved performance on each task compared to single-task learning (STL) baselines.

This chapter offers a comprehensive perspective by combining these two directions (tabular feature integration and multi-task learning) on the potential of combining data modality (textual vs. structured) and learning paradigms (single-task vs. multi-task) in advancing Arabic sentiment analysis.

## 5.2 Experimental analysis of Enhancing Arabic Sentiment Analysis through Multimodal-Toolkit

In this work, we conducted a series of comprehensive experiments to evaluate the effectiveness of integrating tabular features with textual data using the ArsenTD-Lev dataset. The primary objective was to assess whether the incorporation of structured information could enhance the performance of models in dialectal (Levantine dialect) Arabic sentiment analysis. To this end, we systematically compared our proposed feature combination approach against several robust baseline models widely used in the sentiment analysis literature.

### 5.2.1 Experimental Setting

Each feature combination strategy presented in Table 4.2 was empirically evaluated through experimentation. The model was trained to perform Arabic sentiment classification by optimizing the cross-entropy loss function. Training was conducted over 3 to 5 epochs using a learning rate of 2e-5 and a batch size of 32. A summary of the experimental results is provided in Table 5.1.

### 5.2.2 Evaluation Metrics

To assess the performance of the model, we employed the F1-score, a widely used evaluation metric that harmonizes precision and recall into a single measure. As defined in Equation 5.1, the F1-score provides a balanced representation of the model's accuracy, particularly in cases of class imbalance.

$$F1 - Score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \tag{5.1}$$

Precision quantifies the proportion of correctly identified positive instances among all instances predicted as positive, whereas recall measures the model's capacity to retrieve all actual positive instances within the dataset.

### 5.2.3 Compared Baselines

To provide a comprehensive evaluation of the proposed feature combination methods, we benchmarked their performance against several widely adopted baseline models in sentiment analysis. These models include:

XLM-T: A multilingual transformer model specifically optimized for social media text, particularly Twitter. XLM-T demonstrates robust performance in multilingual contexts, making it particularly suitable for sentiment classification, text categorization, and language identification tasks involving diverse linguistic data [153].

GigaBERT: An extensively scaled-up version of the BERT architecture [154], GigaBERT is designed to process vast corpora with greater depth. Its increased capacity enables the learning of intricate linguistic representations, thereby achieving superior performance across a range of natural language understanding tasks [155].

AraBERT: A transformer-based model tailored specifically for Modern Standard Arabic (MSA) and dialectal variants. AraBERT has consistently outperformed multilingual baselines on a range of Arabic NLP benchmarks, establishing itself as a foundational model for Arabic language processing tasks [155].

Arabic-ALBERT: A lightweight adaptation of BERT for the Arabic language, Arabic-ALBERT retains high performance while significantly reducing computational complexity. Its efficiency makes it particularly advantageous for scenarios requiring lower resource consumption without compromising linguistic accuracy [156].

**Table 5.1:** Comparison of combining methods with results on the baseline transfomers using F1 score metric, the best performing model is in bold

| **Model** | ***Text only*** | ***Concat*** | ***Gating*** | ***Weighted Sum*** |
|---|---|---|---|---|
| XLM-T | **56.50** | **63.25** | **59.04** | **66.29** |
| Giga BERT V4 | 48.17 | 59.00 | 52.00 | 58.00 |
| Arabertv0.1 | 54.57 | 42.85 | 54.95 | 54.49 |
| Arabertv0.2-base | 54.26 | 55.60 | 49.76 | 57.22 |
| Arabertv0.2-large | 54.01 | 41.39 | 58.90 | 59.83 |
| Arabic-ALBERT-base | 48.47 | 51.61 | 40.87 | 50.83 |
| Arabic-ALBERT-large | 45.59 | 59.49 | 51.30 | 61.65 |

### 5.2.4 Experimental Results and discussions

As presented in Table 5.1, XLM-T consistently outperforms other models in Arabic sentiment classification, achieving the highest F1-scores across various feature combination strategies. This superior performance is largely attributable to the architecture of XLM-T, which is well-suited for capturing semantic relationships between text pairs. Among the evaluated fusion strategies, the Weighted Sum approach yielded the best results, surpassing Concat, Gating, and text-only combining methods. AraBERT-based models displayed variable performance, while Arabic-ALBERT models generally exhibited lower effectiveness. These findings confirm that the integration of multimodal data significantly enhances Arabic sentiment analysis performance compared to using only textual data.

## 5.3 Experimental analysis of Multi-task learning for multi-dialect Arabic sentiment classification and sarcasm detection

In this study, both single-task and multi-task learning (MTL) paradigms were examined using state-of-the-art pre-trained Arabic language models, including AraBERT, MARBERT, and AraELECTRA. MTL architecture was explored by leveraging diverse datasets and pre-trained models during the training phase, aiming to evaluate its effectiveness in enhancing performance across related tasks such as sentiment classification and sarcasm detection.

### 5.3.1 Datasets

Data collection is a critical component of machine learning, often requiring substantial time and resources, with no guaranteed assurance of its direct relevance to the intended task. In this study, we utilize three benchmark datasets that are integral to the development and evaluation of the proposed framework: ArSarcasm, ArSentD-Lev, NADI, and ASTD datasets.

ArSentD-Lev is a sentiment-labeled Arabic Twitter dataset focused on the Levantine dialect [108]. It comprises 4,000 tweets annotated for sentiment, topic, and sentiment target. The dataset presents significant linguistic and contextual challenges due to its coverage of diverse domains such as politics, religion, sports, entertainment, and personal topics. Tweets were systematically collected to ensure balanced representation across four Levantine countries: Syria, Jordan, Lebanon, and Palestine. A detailed summary of the dataset's characteristics is provided in Table 5.2.

**Table 5.2:** Details of the different annotations of the ArSentD-Lev dataset.

| **Topic** | | **Sentiment** | | **Expression** | |
|---|---|---|---|---|---|
| Sports | 12.12% | Positive | 20.1% | Explicit | 73.6% |
| Religions | 9.83% | Very positive | 10.7% | Implicit | 4.3% |
| Personal | 32.6% | Very negative | 16.3% | None | 22.1% |
| Politics | 37.63% | Negative | 30.8% | | |
| Entertainment | 4.35% | Neutral | 22.13% | | |

The Arabic Sentiment Tweets Dataset (ASTD) [97] is a substantial annotated corpus comprising 10,006 tweets primarily sourced from Egyptian Arabic. The tweets are categorized into four sentiment classes: Positive, Negative, Neutral, and Mixed. This dataset is particularly valuable for advancing Arabic sentiment analysis. It captures the nuanced linguistic expressions and sentiment cues present in dialectal Arabic across a wide array of topics.

The ArSarcasm dataset [149] consists of approximately 10,547 manually annotated Arabic tweets, curated for the dual tasks of sarcasm detection and sentiment analysis. Annotations were performed by native arabic speakers, with sarcasm labels divided into sarcastic and non-sarcastic categories, wheras sentiment labels classified as positive, negative, or neutral. This dataset was constructed by re-annotating entries from previously established corpora, including ASTD [97] and SemEval 2017 [157]. Additionally, it includes dialect annotations to enhance its utility in dialect-specific analysis. A detailed summary of the dataset's structure and annotation schema is presented in Table 5.3.

**Table 5.3:** Details of the different annotations of the ArSarcasm dataset.

| | **ArSarcasm**$_{sarcasm}$ | | **ArSarcasm**$_{senti}$ | | | |
|---|---|---|---|---|---|---|
| **Dialect** | **Non-Sarcastic** | **Sarcastic** | **Negative** | **Neutral** | **Positive** | **Total** |
| MSA | 6,431 | 631 | 1,893 | 42,01 | 968 | 7,062 |
| Levantine | 433 | 118 | 239 | 178 | 134 | 551 |
| Egyptian | 1,584 | 799 | 1,179 | 733 | 471 | 2,383 |
| Gulf | 397 | 122 | 200 | 218 | 101 | 519 |
| Maghrebi | 20 | 12 | 18 | 10 | 4 | 32 |
| Total | 8,865 | 1,682 | 3,529 | 5,340 | 1,678 | 10,547 |

The Nuanced Arabic Dialect Identification (NADI) dataset [158] comprises a large collection of Arabic tweets annotated with arabic Country and Provinces labels. Developed as part of the NADI shared tasks, this dataset serves as a pivotal resource for advancing research in Arabic dialect identification. It spans tweets from 21 Arab countries and 100 provinces, rendering it the most extensive and granular Arabic dialect corpus to date. For the purposes of this study, we labeled the dataset with the respective dialects to support dialect identification experiments. A detailed overview of the NADI dataset is provided in Table 5.4.

**Table 5.4:** Details of the different annotations of the NADI dataset.

| Country Name | Provinces | Total | % | Dialect |
|---|---|---|---|---|
| Algeria | 7 | 2,214 | 7.15 | Maghrebi |
| Bahrain | 1 | 238 | 0.77 | Gulf |
| Djibouti | 1 | 271 | 0.88 | Djiboutian |
| Egypt | 21 | 6,635 | 21.43 | Egyptian |
| Iraq | 12 | 3,816 | 12.33 | Levantine |
| Jordan | 2 | 634 | 2.05 | Levantine |
| Kuwait | 2 | 592 | 1.91 | Gulf |
| Lebanon | 3 | 905 | 2.92 | Levantine |
| Libya | 5 | 1,6 | 5.17 | Maghrebi |
| Mauritania | 1 | 255 | 0.82 | Maghrebi |
| Morocco | 5 | 1,579 | 5.10 | Maghrebi |
| Oman | 6 | 1,615 | 5.22 | Gulf |
| Palestine | 2 | 624 | 2.02 | Levantine |
| Qatar | 2 | 399 | 1.29 | Gulf |
| Saudi Arabia | 10 | 3,455 | 11.16 | Gulf |
| Somalia | 1 | 312 | 1.01 | Somalian |
| Sudan | 1 | 312 | 1.01 | Maghrebi |
| Syria | 5 | 1,595 | 5.15 | Levantine |
| Tunisia | 4 | 1,122 | 3.62 | Maghrebi |
| UAE | 5 | 1,548 | 5.00 | Gulf |
| Yemen | 4 | 1,236 | 3.99 | Gulf |
| Total | 100 | 30,957 | 100 | / |

### 5.3.2 Data-Preprocessing

Pre-processing refers to the systematic transformation of raw data into a structured and analyzable format suitable for machine learning algorithms. This step is often indispensable in data mining and analysis pipelines, as it significantly enhances the quality and accuracy of model outputs by reducing noise, standardizing input formats, and ensuring data consistency [159].

In this framework, we employed the Python module "Arabert_preprocessing",

which offers a suite of functions specifically designed to preprocess Arabic dialectal text for compatibility with AraBERT-based models. This module performs essential linguistic normalization tasks, including diacritic removal, character normalization, word segmentation, and sentence tokenization. It integrates the Farasa library alongside the BertWordPieceTokenizer to generate input tensors optimized for AraBERT encoders. These preprocessing steps are critical in mitigating the morphological and orthographic complexity of Arabic, thereby enhancing the downstream performance of sentiment analysis models [160].

### 5.3.3 Evaluation Metrics

We measures the performance of the recommended framework using the following standard indicators.

a. Accuracy is a common evaluation metric for classification models. As given in Eq. 5.2 it measures the ratio of correct predictions to the total number of predictions.

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \tag{5.2}$$

b. A useful measure called F1-Score that can evaluate two classifiers. It combines both recall and precision into one metric as given in Eq. 5.3.

$$F1 - Score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \tag{5.3}$$

Where Precision measures how often the model correctly identifies the positive class. It refers to the ratio of true positive cases out of all predicted positive cases, which includes false positives (FP) and true positives (TP), as stated in Eq. 5.4.

$$Precision = \frac{TP}{(FP + TP)} \tag{5.4}$$

Recall reflects how well a model identifies all the positive instances in a dataset. It is also called sensitivity. The calculation involves dividing the count of true positives (TP) by the total of true positives (TP) and false negatives (FN), as given in Eq. 5.5.

$$Recall = \frac{TP}{(FN + TP)} \tag{5.5}$$

### 5.3.4 Experimental settings

The experiments were conducted using Google Colab Pro [161], equipped with an NVIDIA A100 GPU (40 GB VRAM), 32 GB of RAM, and approximately 100 GB of storage. Model implementation and training were carried out using the PyTorch deep learning framework [162]. The dataset was divided into three subsets: 80% for training, 10% for validation, and 10% for testing. The evaluation set helps in adjusting parameters and making iterative improvements. While the test set provides an unbiased assessment of the model's final capabilities. By using 80% of the data for training, we ensure the model has ample exposure to a wide range of examples, which enhances its ability to generalize to new, unseen data. We also fine-tuned the Multi-Task Learning (MTL) model by optimizing both shared and task-specific layers. This was achieved through the careful selection of key hyperparameters. All transformer-based models were configured with a maximum input sequence length of 512 tokens. The learning rate was set to $2 \times 10^{-5}$ and optimized using the Adam optimizer [163]. Batch sizes ranged between 16 and 32, depending on the model and GPU memory constraints. A dropout rate of 0.1 was applied to mitigate overfitting. Each model underwent end-to-end training for an average of 5 to 10 epochs, ensuring robust convergence and improved generalization performance.

### 5.3.5 Comparison details

To evaluate our findings, we developed distinct baseline models for each task in one scenario for single-task model comparison, and additional comparisons that integrate all tasks in other scenarios for multi-task learning model assessment.

#### Single task models

In this experiment, we compare the performance of transformer models, including AraBERT, MARBERT, and ARAELECTRA, on the specified datasets as single-task learning models.

#### MTL models

Experiments shown in Tables 2.6, 2.7, 2.8, and 2.9 illustrate the performance of our proposed models in comparison to baseline models, based on tasks (binary, ternary, and quaternary combinations) as MTL models on the specified datasets.

## 5.4 Results and discussion

We organized our experiments into four distinct series. In Experimental Series 1, we evaluated single-task learning (STL) models to establish baseline performance. Experimental Series 2 investigated the efficacy of multi-task learning (MTL) by exploring binary, ternary, and quaternary task configurations, as detailed in Tables 2.6 to 2.8. This series specifically aimed to assess the influence of Arabic sarcasm detection on sentiment classification performance. Furthermore, we

integrated Arabic dialect identification as an auxiliary task to enhance the overall performance of the proposed models on both sentiment and sarcasm detection tasks. In Experimental Series 3, we benchmarked our top-performing model, as presented in Table 2.9, which achieved the highest F1-score and precision across all evaluated datasets, against state-of-the-art methods reported in the existing literature. Finally, in the last experimental series, we assessed the generalizability of our best-performing models by applying them to a distinct dataset ASTD, as shown in Table 2.10—to further validate and reinforce the robustness of our findings. The configuration details of the evaluated model are as follows: MARBERT, Arabertv2-large, and ARAELECTRA: are a single task models fine-tuned on all four datasets. Other models: are Multi-Task Learning models with Arabert v2-large utilized in the shared part and the mentioned datasets on the specific-task part. The details are shown in Table 5.5.

**Table 5.5:** Setup details of our MTL models.

| Type | Model | Shared-Part | Task-specific Part | |
|---|---|---|---|---|
| Binary | MTLBinary1 | Arabertv2-large | $\text{ArSarcasm}_{senti}$ | Sentiment classification |
| | | | NADI | Dialect detection |
| | MTLBinary2 | Arabertv2-large | ArSentD-Lev | Sentiment classification |
| | | | NADI | Dialect detection |
| | MTLBinary3 | Arabertv2-large | $\text{ArSarcasm}_{senti}$ | Sentiment classification |
| | | | $\text{ArSarcasm}_{sarcasm}$ | Sarcasm detection |
| | MTLBinary4 | Arabertv2-large | ArSentD-Lev | Sentiment classification |
| | | | $\text{ArSarcasm}_{sarcasm}$ | Sarcasm detection |
| Ternary | MTLTernary1 | Arabertv2-large | $\text{ArSarcasm}_{senti}$ | Sentiment classification |
| | | | NADI | Dialect detection |
| | | | $\text{ArSarcasm}_{sarcasm}$ | Sarcasm detection |
| | MTLTernary2 | Arabertv2-large | ArSentD-Lev | Sentiment classification |
| | | | NADI | Dialect detection |
| | | | $\text{ArSarcasm}_{sarcasm}$ | Sarcasm detection |
| | MTLTernary3 | Arabertv2-large | ArSentD-Lev | Sentiment classification |
| | | | $\text{ArSarcasm}_{senti}$ | Sentiment classification |
| | | | NADI | Dialect detection |
| | MTLTernary4 | Arabertv2-large | $\text{ArSarcasm}_{sarcasm}$ | Sarcasm detection |
| | | | $\text{ArSarcasm}_{senti}$ | Sentiment classification |
| | | | ArSentD-Lev | Sentiment classification |
| Quaternary | MTLQuaternary | Arabertv2-large | $\text{ArSarcasm}_{sarcasm}$ | Sarcasm detection |
| | | | $\text{ArSarcasm}_{senti}$ | Sentiment classification |
| | | | ArSentD-Lev | Sentiment classification |
| | | | NADI | Dialect detection |

### 5.4.1 Experimental series 1

In this scenario, we evaluated single-task learning baseline models across the aforementioned datasets to identify the model that consistently yields the highest

performance for integration into our proposed methodology. As shown in Table 5.6, AraBERTv2-large demonstrated superior accuracy across all evaluated datasets when compared to other baseline models. The only exception was observed in the ArSarcasm$_{sarcasm}$ (sarcasm specific-task), where ARAELECTRA outperformed AraBERTv2-large. This result may be attributed to the ARAELECTRA's smaller size and reduced number of parameters, which contribute to faster training times and a lower risk of overfitting.

**Table 5.6:** Evaluation of single-task learning models. Best results are shown in bold.

| | **ArSarcasm$_{senti}$** | | **ArSarcasm$_{sarcasm}$** | | **ArSentD-Lev** | | **NADI** | |
|---|---|---|---|---|---|---|---|---|
| | **SC** | | **SD** | | **SC** | | **DI** | |
| Model | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) |
| MARBERT | 74.88 | 70.81 | 85.21 | 72.52 | 55.25 | 51.65 | 82.00 | 82.00 |
| Arabertv2-large | **75.83** | **72.89** | 85.97 | 73.29 | **58.75** | **57.50** | **83.00** | **82.93** |
| ARAELECTRA | 75.36 | 70.62 | **86.07** | **73.92** | 58.00 | 45.93 | 77.75 | 77.41 |

### 5.4.2 Experimental series 2

In this scenario, we conducted experiments with several multi-task learning (MTL) models, utilizing AraBERTv2-large as the shared encoder across tasks. Table 5.7 presents the results for four different task pair combinations. The table shows that the MTLBinary3 and MTLBinary4 configurations achieved the highest performance on both sentiment classification (SC) and sarcasm detection (SD) tasks. This improved performance can be attributed to the suitability of the ArSarcasm and ArSentD_Lev datasets.

Table 5.8 presents the results obtained from four configurations involving the Ternary of task combinations. Among these, the MTLTernary3 model demonstrates the highest performance, achieving an accuracy of 77.06% on the ArSarcasm$_{senti}$ dataset and 61.50% on the ArSentD-Lev dataset. This performance gain is primarily attributed to the dialectal variation correction facilitated by the inclusion of the NADI dataset (dialect identification specific task). The dialectal differences between ArSentD_Lev and ArSarcasm$_{senti}$, compounded by the constrained lexical coverage of existing Arabic language models, are effectively mitigated through this auxiliary task. Consequently, the model benefits from enhanced generalization and improved cross-dialectal sentiment classification.

The MTLTernary4 model reported the highest F1-score across both ArSarcasm$_{senti}$ and ArSentD_Lev, achieving 73.44% and 58.67%, respectively.

Additionally, it exhibited superior performance on the ArSarcasm$_{sarcasm}$ task, with an accuracy of 86.16% and an F1-score of 74.53%. These results affirm the model's ability to benefit from shared learning signals across interrelated sentiment and sarcasm tasks.

Further, Table 5.9 presents the results of the MTLQuaternary model, which integrates all four tasks. This configuration achieved an F1-score of 72.94% on the ArSarcasm$_{senti}$ task, outperforming single-task models such as ARAELECTRA and AraBERTv2, as previously shown in Table 5.6. On the ArSentD-Lev dataset, MTLQuaternary also surpassed the ARAELECTRA baseline, obtaining an F1-score of 54.61%.

These results validate the efficacy of the proposed MTL approach in leveraging both global and task-specific contextual representations via shared and task-specific parts. By learning across multiple datasets and tasks, the MTL models significantly enhance generalization capability, classification accuracy, and computational efficiency compared to their single-task models.

**Table 5.7:** Performance of Binary-task learning models. Best results are shown in bold.

| | **ArSarcasm$_{senti}$** | | **ArSarcasm$_{sarcasm}$** | | **ArSentD-Lev** | | **NADI** | |
|---|---|---|---|---|---|---|---|---|
| | **SC** | | **SD** | | **SC** | | **DI** | |
| Model | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) |
| MTLBinary1 | 76.59 | 73.15 | - | - | - | - | 82.75 | 82.70 |
| MTLBinary2 | - | - | - | - | 61.25 | 57.51 | 82.75 | 82.69 |
| MTLBinary3 | **77.63** | **73.96** | 86.26 | 76.39 | - | - | - | - |
| MTLBinary4 | - | - | **87.77** | **76.42** | **61.75** | **59.46** | - | - |

**Table 5.8:** Performance of Ternary-task learning models. Best results are shown in bold.

| Model | **ArSarcasm$_{senti}$** | | **ArSarcasm$_{sarcasm}$** | | **ArSentD-Lev** | | **NADI** | |
|---|---|---|---|---|---|---|---|---|
| | **SC** | | **SD** | | **SC** | | **DI** | |
| | **Acc (%)** | **F1 (%)** | **Acc (%)** | **F1 (%)** | **Acc (%)** | **F1 (%)** | **Acc (%)** | **F1 (%)** |
| MTLTernary1 | 74.98 | 72.78 | 85.59 | 73.35 | - | - | 61.50 | 59.25 |
| MTLTernary2 | - | - | 85.88 | 72.34 | 59.69 | 55.78 | 59.25 | 56.23 |
| MTLTernary3 | **77.06** | 73.14 | - | - | **61.50** | 56.68 | **82.75** | **82.71** |
| MTLTernary4 | 76.21 | **73.44** | **86.16** | **74.53** | 60.31 | **58.67** | - | - |

**Table 5.9:** Performance of quaternary-task learning models

| Model | **ArSarcasm$_{senti}$** | | **ArSarcasm$_{sarcasm}$** | | **ArSentD-Lev** | | **NADI** | |
|---|---|---|---|---|---|---|---|---|
| | **SC** | | **SD** | | **SC** | | **DI** | |
| | **Acc (%)** | **F1 (%)** | **Acc (%)** | **F1 (%)** | **Acc (%)** | **F1 (%)** | **Acc (%)** | **F1 (%)** |
| MTLQuaternary | 72.94 | 72.94 | 87.01 | 74.55 | 57.25 | 54.61 | 61.50 | 59.63 |

### 5.4.3 Experimental series 3

This study illustrates our top-performing model, distinguished by its superior F1-score and precision across all employed datasets, with contemporary state-of-the-art models.

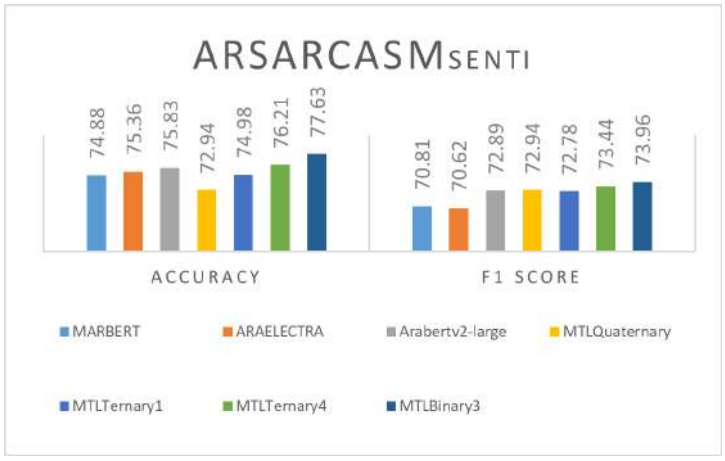
As presented in Table 5.10, the multi-task learning (MTL) models consistently outperform the single-task learning (STL) models in both F1-score and accuracy on the ArSarcasm and ArSentD-Lev datasets. Notably, MTL models trained with two dataset-task pairs demonstrated enhanced performance relative to those trained on three or four tasks. This performance differential may be attributed to the effects of data imbalance, which can adversely influence model optimization and generalization across heterogeneous tasks. The findings demonstrate that training a multi-task learning (MTL) model on a unified task type, such as sentiment classification using multiple datasets characterized by diverse dialects, contextual domains, and class

distributions, can significantly enhance the contextual representations learned by pre-trained language models and yield superior outcomes than training on a single task.

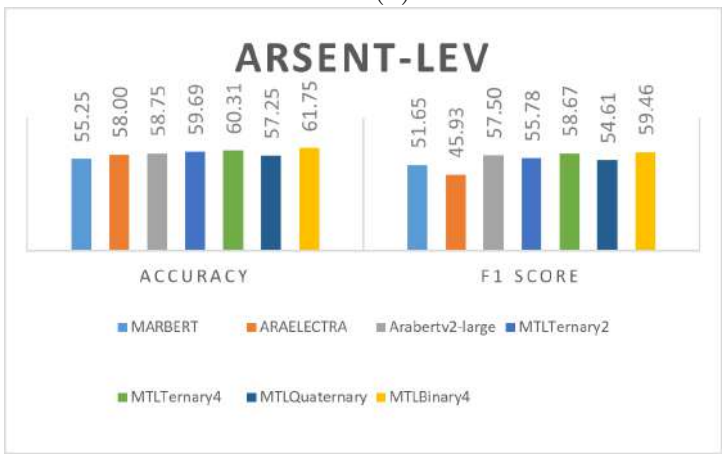
However, on the NADI dataset which is used for dialect identification tasks, the single-task learning (STL) model specifically AraBERTv2 has demonstrated marginally superior performance compared to the multi-task learning (MTL) configurations.

This outcome can be attributed to several factors. One prominent issue is negative transfer or task interference, wherein the joint training of multiple tasks results in mutual degradation of performance rather than helping performance. Additionally, task dominance may occur in MTL settings, whereby tasks that are inherently easier to learn such as sentiment classification or sarcasm detection receive disproportionately more attention from the shared model capacity, thus overshadowing more complex tasks like dialect identification. Another critical factor is label granularity. Dialect identification involves fine-grained classification across a large number of closely related dialects, whereas sentiment and sarcasm detection tasks often rely on more coarse-grained labels (e.g., positive/negative/neutral/very positive/very negative or sarcastic/non-sarcastic). This disparity may lead to representation dilution, wherein the shared parameters in the MTL architecture become biased toward learning more generalized representations that are less effective for subtler distinctions required in dialect classification. In contrast, a single-task learning (STL) model such as AraBERTv2, when trained exclusively on the dialect identification task, retains its full model capacity for learning the fine-grained linguistic distinctions necessary for accurate classification. This dedicated focus allows the model to capture subtle phonological, morphological, and lexical variations across dialects, thereby yielding superior performance relative to MTL configurations.

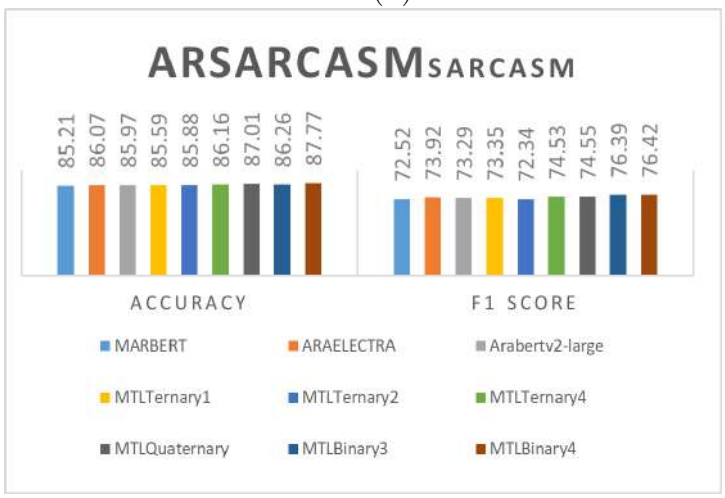
Nevertheless, our findings do not suggest that multi-task learning (MTL) is universally optimal across all tasks. Instead, underscore its effectiveness in enhancing sentiment classification, particularly through the shared representations derived from related tasks such as sarcasm detection, which mutually reinforce each other's performance. Figure 5.1 demonstrates the impact of incorporating sarcasm detection as an auxiliary task on sentiment classification within multi-task learning (MTL) models, compared to single-task learning (STL) models. The results clearly show that most MTL models outperform STL models in sentiment classification, attaining an F1-score of 73.96% and an accuracy of 77.63% on the ArSarcasm*senti* dataset, and achieving an F1-score of 59.46% and an accuracy of 61.75% on the ArSentD-Lev dataset. Moreover, for sarcasm detection, the top-performing MTL model significantly surpasses all other models, registering an F1-score of 76.42% and an accuracy of 87.77% on the ArSarcasm*sarcasm* dataset.




(a)




(b)




(c)

**Figure 5.1:** Experimental results of sarcasm detection and Arabic sentiment classification tasks using different variety of models. (a) Evaluation of the models on ArSarcasm$_{senti}$ dataset. (b) Evaluation of the models on ArSentD-Lev dataset. (c) Evaluation of the models on ArSarcasm$_{sarcasm}$ dataset.

**Table 5.10:** Comparison of our best models with the state-of-the-art models.

| Model | **ArSarcasm$_{senti}$** | | **ArSarcasm$_{sarcasm}$** | | **ArSentD-Lev** | | **NADI** | |
|---|---|---|---|---|---|---|---|---|
| | **SC** | | **SD** | | **SC** | | **DI** | |
| | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) | **Acc** (%) | **F1** (%) |
| MARBERT [26] | - | 71.50 | - | 76.30 | 60.38 | - | - | - |
| ARAELECTRA[58] | - | - | - | - | - | 57.20 | - | - |
| FreezeArabic-BERT mini[164] | - | - | 83.98 | - | - | - | - | - |
| AraBERT[57] | - | - | - | - | 59.40 | - | - | - |
| hULMonA[165] | - | - | - | - | - | 51.10 | - | - |
| AraBERT with Ar-PuFi Training[166] | - | - | - | 76.00 | - | - | - | - |
| Arabertv2 | 75.83 | 72.89 | 85.97 | 73.29 | 58.75 | 57.50 | **83.00** | **82.93** |
| ARAELECTRA | 75.36 | 70.62 | 86.07 | 73.92 | 58.00 | 45.93 | 77.75 | 77.41 |
| MTLBinary3 | **77.63** | **73.96** | 86.26 | 76.39 | - | - | - | - |
| MTLBinary4 | - | - | **87.77** | **76.42** | **61.75** | **59.46** | - | - |
| MTLBinary1 | 76.59 | 73.15 | - | - | - | - | 82.75 | 82.70 |
| MTLBinary2 | - | - | - | - | 61.25 | 57.51 | 82.75 | 82.69 |
| MTLTernary3 | 77.06 | 73.14 | - | - | 61.50 | 56.68 | 82.75 | 82.71 |
| MTLTernary4 | 76.20 | 73.44 | 86.16 | 74.53 | 60.31 | 58.67 | - | - |
| MTLQuaternary | 72.94 | 72.94 | 87.01 | 74.55 | 57.25 | 54.61 | 61.50 | 59.63 |

### 5.4.4 Experimental series 4

To strengthen the conclusions of this study, our best performing models for sentiment classification and sarcasm detection were additionally evaluated on an external benchmark dataset, the Arabic Sentiment Tweets Dataset (ASTD) [97]. Table 5.11 summarizes the performance results of the binary-task learning models on ASTD. The results indicate that multi-task learning (MTL) models consistently outperform single-task learning (STL) models with respect to both F1-score and accuracy across the evaluated dataset. This enhancement is primarily attributed to the capacity of MTL models to mitigate overfitting during training by enabling the model to generalize better through shared learning. Furthermore, the incorporation of sarcasm detection significantly contributed to improved sentiment classification performance within the MTL framework, and vice versa, highlighting the mutual benefits of task synergy.

**Table 5.11:** Performance of Binary-task learning models on ASTD dataset. Best results are shown in bold.

| Model | **ASTD** | |
|---|---|---|
| | **Acc** (%) | **F1** (%) |
| mBERT [154] | - | 46.30 |
| AraBERT [57] | - | 57.50 |
| Arabertv2 | 71.80 | 56.46 |
| MTLBinary3 | 72.00 | 57.77 |
| MTLBinary4 | **75.00** | **59.07** |

## 5.5 Conclusion

This chapter presented an extensive evaluation of two complementary approaches aimed at improving Arabic sentiment analysis: the integration of tabular features with textual inputs (Enhancing Arabic Sentiment Analysis through Multimodal-Toolkit), and the application of multi-task learning (MTL) for (Multi-Dialect Arabic Sentiment classification and Sarcasm Detection). The experimental findings reveal several key insights regarding the effectiveness of these strategies in advancing performance across dialectal Arabic datasets.

First, the incorporation of tabular features such, as numerical and categorical data, demonstrated clear benefits when combined with transformer-based textual encoders. Notably, the XLM-T model, when paired with the Weighted Sum combination method, consistently achieved superior results, indicating that well-designed feature integration can enrich semantic representations and mitigate

limitations associated with text-only modeling. This finding emphasizes the importance of multimodal learning in capturing nuanced linguistic patterns within Arabic dialects, particularly the Levantine dialect.

Second, the MTL experiments confirmed that learning multiple related tasks simultaneously can lead to improved generalization and performance, particularly in sentiment classification and sarcasm detection. The shared representation space facilitated knowledge transfer between tasks, allowing the model to learn more robust and context-aware features. However, challenges such as task interference and representational imbalance were also observed, especially in the dialect identification task, suggesting that task compatibility and data characteristics are critical factors in MTL design.

In conclusion, both studies highlight the significance of enriched input representations and multi-objective learning frameworks in advancing the effectiveness of pre-trained language models for Arabic sentiment analysis. While each approach contributes unique strengths, their integration presents a promising direction for future research. Specifically, incorporating structured tabular features within multi-task learning architectures may further optimize model performance across diverse and linguistically rich Arabic datasets.

# Chapter 6

# General Conclusion and Future Work

## 6.1 General Conclusion

This dissertation investigated the complexities of Arabic sentiment analysis and proposed advanced deep learning approaches to improve classification performance in the presence of dialectal variation, morphological richness, and contextual ambiguity. The research was grounded in a comprehensive exploration of sentiment analysis methodologies, from traditional lexicon-based and machine learning approaches to modern transfer learning frameworks, particularly those employing transformer-based architectures. The focus on Arabic language processing highlighted the challenges unique to the language such as dialectal fragmentation, scarcity of annotated datasets, and the presence of sarcasm and demonstrated the need for innovative solutions tailored to these constraints.

Building on this foundation, two new methodologies were proposed and evaluated. The first introduced a multimodal architecture that integrates textual, categorical, and numerical features using the ArsenTD-Lev dataset. This approach demonstrated the effectiveness of combining tabular data to capture nuanced sentiment patterns, particularly in dialectal contexts. The second contribution employed a Multi-Task Learning (MTL) framework for simultaneous sentiment classification, sarcasm detection, and dialect identification. This architecture capitalized on shared linguistic representations to improve generalization across tasks, confirming the interrelated nature of sentiment and sarcasm in Arabic textual data.

The experimental analyses conducted in this study provided empirical validation of both methodologies. Results indicated that multimodal integration significantly improves sentiment classification performance compared to using text-only method. Additionally, the MTL framework outperformed single-task models by leveraging task interdependencies, although certain challenges such as task interference and imbalanced data representation were observed. These findings underscore the value of transfer learning paradigms and task-specific tuning in achieving robust Arabic NLP systems.

This research contributes to the growing body of work in low-resource language

processing by advancing methods that enhance model adaptability, generalization, and contextual understanding. It lays the groundwork for future studies aiming to scale Arabic sentiment analysis across dialects and tasks, providing methodological insights and architectural innovations that can be extended to other under-resourced languages.

## 6.2 Future Work

While this dissertation presents significant advancements in Arabic sentiment analysis, several avenues remain open for future exploration and enhancement. First, the multimodal architecture developed in this work can be further extended to include additional modalities such as audio, image, or video content with different datasets. These media forms are increasingly prevalent in social media and can provide essential context for understanding sentiment, emotion, or sarcasm in user-generated content. Extending the model to handle these diverse data types would broaden its applicability to tasks such as multimodal emotion recognition, fake news detection, and cross-modal sentiment analysis.

Second, although the current Multi-Task Learning (MTL) framework demonstrates improved performance across sentiment classification and sarcasm detection further research is needed to mitigate issues related to task interference and negative transfer. Future work could investigate adaptive weighting schemes or hierarchical task modeling mechanisms to enhance cooperation among tasks.

A promising direction lies in the integration of structured tabular features within the MTL framework. While this study separately evaluated multimodal learning using tabular features and MTL using textual features, a unified model that incorporates both could leverage the strengths of each modality. By enabling shared and task-specific representations that jointly model text and tabular data, such a system could better capture complex patterns across dialects and sentiment expressions.

Additionally, future work should explore model interpretability and explainability, particularly in real-world applications. As sentiment analysis systems are increasingly used in domains such as healthcare, politics, and business intelligence, understanding model decisions becomes essential. Methods such as attention heatmaps and counterfactual examples tools adapted to Arabic NLP can increase transparency and user trust in the deployed systems.

In conclusion, the integration of tabular features within a multi-task learning framework presents an innovative frontier for Arabic sentiment analysis. Coupled with advances in data augmentation, task optimization, and interpretability, these future directions have the potential to build more robust, generalizable, and contextually aware NLP systems for Arabic and other low-resource languages.

# Bibliography


[1] Allen Rubi and Earl R Babbie. *Empowerment series: Research methods for social work*. 2016.

[2] David L Olson, Dursun Delen, and Yanyan Meng. "Comparative analysis of data mining methods for bankruptcy prediction". In: *Decision Support Systems* 52.2 (2012), pp. 464–473.

[3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes twitter users: real-time event detection by social sensors". In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 851–860.

[4] Khurshid Ahmad, David Cheng, and Yousif Almas. "Multi-lingual sentiment analysis of financial news streams". In: *Proc. of the 1st Intl. Conf. on Grid in Finance*. 2006.

[5] Yousif Almas and Khurshid Ahmad. "A note on extracting 'sentiments' in financial news in English, Arabic & Urdu". In: *The Second Workshop on Computational Approaches to Arabic Script-based Languages*. 2007, pp. 1–12.

[6] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. "Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques". In: *Third IEEE International Conference on Data Mining*. 2003, pp. 427–434.

[7] Ghadeer Al-Sukkar, Ibrahim Aljarah, and Hamad Alsawalqah. "Enhancing the Arabic Sentiment Analysis Using Different Preprocessing Operators". In: Apr. 2017.

[8] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. URL: https://aclanthology.org/N18-1202/.

[9] Shijie Chen, Yu Zhang, and Qiang Yang. "Multi-task learning in natural language processing: An overview". In: *ACM Computing Surveys* 56.12 (2024), pp. 1–32.

[10] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. "New Avenues in Opinion Mining and Sentiment Analysis". In: *IEEE Intelligent Systems* 28.2 (2013), pp. 15–21.

[11] Wesam Alsabban. “Exploring Sentiment Analysis on Arabic Tweets about the COIVD-19 Vaccines”. In: *Tehnički glasnik* 16 (May 2022), pp. 268–272.

[12] Dhekra Najar and Slim Mesfar. “Opinion mining and sentiment analysis for Arabic on-line texts: application on the political domain”. In: *International Journal of Speech Technology* 20 (Sept. 2017), pp. 1–11.

[13] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. “Lexicon-Based Methods for Sentiment Analysis”. In: *Computational Linguistics* 37.2 (2011), pp. 267–307.

[14] Fazel Keshtkar and Diana Inkpen. “A BOOTSTRAPPING METHOD FOR EXTRACTING PARAPHRASES OF EMOTION EXPRESSIONS FROM TEXTS”. In: *Computational Intelligence* 29.3 (2013), pp. 417–435. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00458.x`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00458.x`.

[15] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. “New Avenues in Opinion Mining and Sentiment Analysis”. In: *IEEE Intelligent Systems* 28.2 (2013), pp. 15–21.

[16] Ahmed Ali and Samhaa El-Beltagy. “Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study”. In: Mar. 2013.

[17] Gilbert Badaro, Ramy Baly, Hazem Hajj, Wassim El-Hajj, Khaled Bashir Shaban, Nizar Habash, Ahmad Al-Sallab, and Ali Hamdi. “A Survey of Opinion Mining in Arabic: A Comprehensive System Perspective Covering Challenges and Advances in Tools, Resources, Models, Applications, and Visualizations”. In: 18.3 (May 2019). URL: `https://doi.org/10.1145/3295662`.

[18] Xiaobo Zhang and Qingsong Yu. “Hotel reviews sentiment analysis based on word vector clustering”. In: *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*. 2017, pp. 260–264.

[19] O. Chapelle, B. Scholkopf, and A. Zien Eds. “Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]”. In: *IEEE Transactions on Neural Networks* 20.3 (2009), pp. 542–542.

[20] Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. “Aspect Extraction with Automated Prior Knowledge Learning”. In: vol. 1. June 2014.

[21] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. “Sentiment analysis algorithms and applications: A survey”. In: *Ain Shams Engineering Journal* 5.4 (2014), pp. 1093–1113. URL: `https://www.sciencedirect.com/science/article/pii/S2090447914000550`.

[22] Lei Zhang, Shuai Wang, and Bing Liu. “Deep learning for sentiment analysis: A survey”. In: *WIREs Data Mining and Knowledge Discovery* 8.4 (2018), e1253.

[23] S Sharath T and Shubhangi Tandon. “Topic Based Sentiment Analysis Using Deep Learning”. In: *arXiv e-prints* (2017), arXiv–1710.

[24] Muhammad Zubair Asghar, Fazal Masud Kundi, Shakeel Ahmad, Aurangzeb Khan, and Furqan Khan. “T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme”. In: *Expert Systems* 35.1 (2018), e12233.

[25] Mark Abraham Magumba, Peter Nabende, and Ernest Mwebaze. “Ontology boosted deep learning for disease name extraction from Twitter messages”. In: *Journal of Big Data* 5 (2018), pp. 1–19.

[26] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. “ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic”. In: *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)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 7088–7105. URL: https://aclanthology.org/2021.acl-long.551/.

[27] Dario Stojanovski, Gjorgji Strezoski, Gjorgji Madjarov, Ivica Dimitrovski, and Ivan Chorbev. “Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages”. In: *Multimedia Tools and Applications* 77 (2018), pp. 32213–32242.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[30] Ahmed Derbala Yacoub, Salwa Slim, and Amal Aboutabl. “A Survey of Sentiment Analysis and Sarcasm Detection: Challenges, Techniques, and Trends”. In: *International Journal of Electrical and Computer Engineering Systems* 15.1 (Jan. 2024), pp. 69–78. URL: https://ijeces.ferit.hr/index.php/ijeces/article/view/2953.

[31] John Blitzer, Mark Dredze, and Fernando Pereira. “Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Ed. by Annie Zaenen and Antal van den Bosch. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 440–447. URL: https://aclanthology.org/P07-1056/.

[32] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Ed. by David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. URL: https://aclanthology.org/D13-1170/.

[33] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. “A systematic study of the class imbalance problem in convolutional neural networks”. In: *Neural Networks* 106 (Oct. 2018), pp. 249–259. URL: http://dx.doi.org/10.1016/j.neunet.2018.07.011.

[34] Harish Thangaraj, Ananya Chenat, Jaskaran Singh Walia, and Vukosi Marivate. *Cross-lingual transfer of multilingual models on low resource African Languages*. 2024. arXiv: 2409.10965 [cs.CL]. URL: https://arxiv.org/abs/2409.10965.

[35] Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. "A Panoramic Survey of Natural Language Processing in the Arab World". In: *Communications of the ACM* 64 (Mar. 2021).

[36] Ali Farghaly and Khaled Shaalan. "Arabic Natural Language Processing: Challenges and Solutions". In: *ACM Transactions on Asian Language Information Processing* 8.4 (Dec. 2009). URL: https://doi.org/10.1145/1644879.1644881.

[37] Abdulhadi Shoufan and Sumaya Alameri. "Natural Language Processing for Dialectical Arabic: A Survey". In: *Proceedings of the Second Workshop on Arabic Natural Language Processing*. Ed. by Nizar Habash, Stephan Vogel, and Kareem Darwish. Beijing, China: Association for Computational Linguistics, July 2015, pp. 36–48. URL: https://aclanthology.org/W15-3205/.

[38] Naaima Boudad, Rdouan Faizi, Rachid Oulad Haj Thami, and Raddouane Chiheb. "Sentiment analysis in Arabic: A review of the literature". In: *Ain Shams Engineering Journal* 9.4 (2018), pp. 2479–2490. URL: https://www.sciencedirect.com/science/article/pii/S2090447917300862.

[39] *Top 10 Languages Used on the Internet*. URL: https://www.accreditedlanguage.com/technology/top-10-languages-used-on-the-internet/.

[40] Abdelali A., Cowie J, and Soliman H.S. "Arabic information retrieval perspectives.Arabic information retrieval perspectives". In: vol. 1. 2004.

[41] Kees Versteegh, Mushira Eid, Alaa Elgibali, Manfred Woidich, and Andrzej Zaborski. *Encyclopedia of Arabic Language and Linguistics, Volume 3*. Leiden, The Netherlands: Brill, 2007. URL: https://brill.com/view/title/12147.

[42] Yasser Salem, Brian Nolan, and Arnold Hensman. "A generic framework for Arabic to English machine translation of simplex sentences using the Role an". In: 2009. URL: https://api.semanticscholar.org/CorpusID:60685617.

[43] Mohammed Dawood and Mohammed Aal-Hajiahmed. "The Ambiguity of Gender in English -Arabic Translation". In: (Oct. 2022).

[44] Ibrahim Badr, Rabih Zbib, and James Glass. "Syntactic Phrase Reordering for English-to-Arabic Statistical Machine Translation". In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Ed. by Alex Lascarides, Claire Gardent, and Joakim Nivre. Athens, Greece: Association for Computational Linguistics, Mar. 2009, pp. 86–93. URL: https://aclanthology.org/E09-1011/.

[45] Peter F. Abboud, Aman Attieh, Ernest N. McCarus, and Raji M. Rammuny. *Modern Standard Arabic*. CUP, 2023.

[46] Mohammed H Al Aqad. "Syntactic analysis of Arabic adverb's between Arabic and English: X bar theory". In: *International Journal of Language and Linguistics* 1.3 (2013), pp. 70–74.

[47] M Abu Shquier and O Abu Shqeer. “Words ordering and corresponding verb-subject agreements in English-Arabic machine translation: Hybrid-based approach”. In: *International Journal of Soft Computing And Software Engineering (JSCSE)* (2012), pp. 49–60.

[48] A Saleh Alduais. “Simple Sentence Structure of Standard Arabic Language and Standard English Language: A Contrastive Study”. In: *International Journal of Linguistics*. (2012). URL: http://www.macrothink.org/journal/index.php/ijl/article/viewFile/2621/pdf.

[49] Jane Wightwick and Mahmoud Gaafar. *Mastering Arabic Grammar*. Bloomsbury Publishing, 2018.

[50] Karin C Ryding. *A reference grammar of modern standard Arabic*. Cambridge university press, 2005.

[51] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. “On the Linguistic Representational Power of Neural Machine Translation Models”. In: *Computational Linguistics* 46.1 (2020), pp. 1–52. URL: https://aclanthology.org/2020.cl-1.1/.

[52] Leena Lulu and Ashraf Elnagar. “Automatic Arabic Dialect Classification Using Deep Learning Models”. In: *Procedia Computer Science* 142 (2018). Arabic Computational Linguistics, pp. 262–269. URL: https://www.sciencedirect.com/science/article/pii/S1877050918321938.

[53] Reem AlYami and Rabeah AlZaidy. “Arabic Dialect Identification in Social Media”. In: *2020 3rd International Conference on Computer Applications and Information Security (ICCAIS)*. 2020, pp. 1–2.

[54] Soha Ahmed, Michel Pasquier, and Ghassan Qadah. “Key issues in conducting sentiment analysis on Arabic social media text”. In: *2013 9th International conference on innovations in information technology (IIT)*. IEEE. 2013, pp. 72–77.

[55] Tagwa Abd Elatif Mohammed. “Review of sentiment analysis for classification Arabic tweets”. In: *International Journal of Emerging Technology and Advanced Engineering* 6.3 (2016), pp. 47–53.

[56] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/.

[57] Wissam Antoun, Fady Baly, and Hazem Hajj. *AraBERT: Transformer-based Model for Arabic Language Understanding*. 2021. arXiv: 2003.00104 [cs.CL]. URL: https://arxiv.org/abs/2003.00104.

[58] Wissam Antoun, Fady Baly, and Hazem M. Hajj. “AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding”. In: *CoRR* abs/2012.15516 (2020). arXiv: 2012.15516. URL: https://arxiv.org/abs/2012.15516.

[59] Rich Caruana. “Multitask learning”. In: *Machine learning* 28 (1997), pp. 41–75.

[60] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. “Which tasks should be learned together in multi-task learning?” In: *International conference on machine learning*. PMLR. 2020, pp. 9120–9132.

[61] Jonathan Baxter. “A model of inductive bias learning”. In: *Journal of artificial intelligence research* 12 (2000), pp. 149–198.

[62] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. “Google’s multilingual neural machine translation system: Enabling zero-shot translation”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 339–351.

[63] Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. “A multi-lingual multi-task architecture for low-resource sequence labeling”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 799–809.

[64] Sebastian Ruder. “An overview of multi-task learning in deep neural networks”. In: *arXiv preprint arXiv:1706.05098* (2017).

[65] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. “Effective approaches to attention-based neural machine translation”. In: *arXiv preprint arXiv:1508.04025* (2015).

[66] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. “Multi-task learning for multiple language translation”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 1723–1732.

[67] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. “The natural language decathlon: Multitask learning as question answering”. In: *arXiv preprint arXiv:1806.08730* (2018).

[68] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. “Recurrent neural network for text classification with multi-task learning”. In: *arXiv preprint arXiv:1605.05101* (2016).

[69] Ronan Collobert and Jason Weston. “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 160–167.

[70] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. “Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser”. In: *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*. 2015, pp. 845–850.

[71] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. “A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1923–1933. URL: https://aclanthology.org/D17-1206/.

[72] Hui Wan. *Multi-task Learning with Multi-head Attention for Multi-choice Reading Comprehension*. 2020. arXiv: 2003.04992 [cs.CL]. URL: https://arxiv.org/abs/2003.04992.

[73] Yu Zhang and Qiang Yang. "A survey on multi-task learning". In: *IEEE transactions on knowledge and data engineering* 34.12 (2021), pp. 5586–5609.

[74] Alex Kendall, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7482–7491.

[75] Shikun Liu, Edward Johns, and Andrew J Davison. "End-to-end multi-task learning with attention". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1871–1880.

[76] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks". In: *International conference on machine learning*. PMLR. 2018, pp. 794–803.

[77] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. "Gradient surgery for multi-task learning". In: *Advances in neural information processing systems* 33 (2020), pp. 5824–5836.

[78] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. "Conflict-averse gradient descent for multi-task learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18878–18890.

[79] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. "Dynamic task prioritization for multitask learning". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 270–287.

[80] Ozan Sener and Vladlen Koltun. "Multi-task learning as multi-objective optimization". In: *Advances in neural information processing systems* 31 (2018).

[81] Hady ElSahar and Samhaa R El-Beltagy. "Building large arabic multi-domain resources for sentiment analysis". In: *International conference on intelligent text processing and computational linguistics*. Springer. 2015, pp. 23–34.

[82] Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. "Subjectivity and sentiment analysis of modern standard Arabic". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011, pp. 587–591.

[83] Alaa Rahma, Shahira Shaaban Azab, and Ammar Mohammed. "A comprehensive survey on Arabic sarcasm detection: approaches, challenges and future trends". In: *IEEE Access* 11 (2023), pp. 18261–18280.

[84] Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. "A large scale Arabic sentiment lexicon for Arabic opinion mining". In: *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*. 2014, pp. 165–173.

[85] Fawaz HH Mahyoub, Muazzam A Siddiqui, and Mohamed Y Dahab. "Building an Arabic sentiment lexicon using semi-supervised learning". In: *Journal of King Saud University-Computer and Information Sciences* 26.4 (2014), pp. 417–424.

[86] Muhammad Abdul-Mageed and Mona T Diab. “Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis.” In: *LREC*. 2014, pp. 1162–1169.

[87] Mahmoud Al-Ayyoub, Safa Bani Essa, and Izzat Alsmadi. “Lexicon-based sentiment analysis of Arabic tweets”. In: *International Journal of Social Network Mining* 2.2 (2015), pp. 101–114.

[88] Nawaf Abdulla, Roa’a Majdalawi, Salwa Mohammed, Mahmoud Al-Ayyoub, and Mohammed Al-Kabi. “Automatic lexicon construction for arabic sentiment analysis”. In: *2014 International Conference on Future Internet of Things and Cloud*. IEEE. 2014, pp. 547–552.

[89] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. “Sentiment Strength Detection in Short Informal Text”. In: *Journal of the American Society for Information Science and Technology* 61 (Dec. 2010), pp. 2544–2558.

[90] Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Alsalman, and Yousef Al-Ohali. “Sentiment analysis of arabic tweets: Feature engineering and a hybrid approach”. In: *arXiv preprint arXiv:1805.08533* (2018).

[91] M’hamed Mataoui, Omar Zelmati, and Madiha Boumechache. “A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic”. In: *Research in Computing Science* 110.1 (2016), pp. 55–70.

[92] Tareq Al-Moslmi, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. “Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis”. In: *Journal of information science* 44.3 (2018), pp. 345–362.

[93] Ahmad Aloqaily, MALAK Al-Hassan, Kamal Salah, Basima Elshqeirat, Montaha Almashagbah, and P Al Hussein Bin Abdullah. “Sentiment analysis for arabic tweets datasets: Lexicon-based and machine learning approaches”. In: *J. Theor. Appl. Inf. Technol* 98.4 (2020), pp. 612–623.

[94] Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. “SAMAR: Subjectivity and sentiment analysis for Arabic social media”. In: *Computer Speech & Language* 28.1 (2014), pp. 20–37.

[95] Aqil M Azmi and Samah M Alzanin. “Aara’–a system for mining the polarity of Saudi public opinion through e-newspaper comments”. In: *Journal of Information Science* 40.3 (2014), pp. 398–410.

[96] Maher Itani, Chris Roast, and Samir Al-Khayatt. “Corpora for sentiment analysis of Arabic text in social media”. In: *2017 8th international conference on information and communication systems (ICICS)*. IEEE. 2017, pp. 64–69.

[97] 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*. 2015, pp. 2515–2519.

[98] Ahmed Y Al-Obaidi and Venus W Samawi. “Opinion mining: Analysis of comments written in arabic colloquial”. In: *Proceedings of the World Congress on Engineering and Computer Science*. Vol. 1. 2016.

[99] Khaled Mohammad Alomari, Hatem M ElSherif, and Khaled Shaalan. "Arabic tweets sentimental analysis using machine learning". In: *International conference on industrial, engineering and other applications of applied intelligent systems*. Springer. 2017, pp. 602–610.

[100] Hamed Al-Rubaiee, Renxi Qiu, and Dayou Li. "Identifying Mubasher software products through sentiment analysis of Arabic tweets". In: *2016 international conference on industrial informatics and computer systems (ciics)*. IEEE. 2016, pp. 1–6.

[101] Lamia Al-Horaibi and Muhammad Badruddin Khan. "Sentiment analysis of Arabic tweets using text mining techniques". In: *First International Workshop on Pattern Recognition*. Vol. 10011. SPIE. 2016, pp. 288–292.

[102] Mohamed Ali Sghaier and Mounir Zrigui. "Sentiment analysis for Arabic e-commerce websites". In: *2016 International Conference on Engineering & MIS (ICEMIS)*. IEEE. 2016, pp. 1–7.

[103] Ramy Baly, Hazem Hajj, Nizar Habash, Khaled Bashir Shaban, and Wassim El-Hajj. "A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in Arabic". In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 16.4 (2017), pp. 1–21.

[104] Hichem Rahab, Abdelhafid Zitouni, and Mahieddine Djoudi. "Siaac: Sentiment polarity identification on arabic algerian newspaper comments". In: *Applied Computational Intelligence and Mathematical Methods: Computational Methods in Systems and Software 2017, vol. 2*. Springer. 2018, pp. 139–149.

[105] Hala Mulki, Hatem Haddad, Mourad Gridach, and Ismail Babaoğlu. "Tw-star at semeval-2017 task 4: Sentiment classification of arabic tweets". In: *Proceedings of the 11th international workshop on semantic evaluation (SEMEVAL-2017)*. 2017, pp. 664–669.

[106] Mohcine Maghfour and Abdeljalil Elouardighi. "Standard and dialectal Arabic text classification for sentiment analysis". In: *Model and Data Engineering: 8th International Conference, MEDI 2018, Marrakesh, Morocco, October 24–26, 2018, Proceedings 8*. Springer. 2018, pp. 282–291.

[107] Awny A Sayed, Enas Elgeldawi, Alaa M Zaki, and Ahmed R Galal. "Sentiment analysis for Arabic reviews using machine learning classification algorithms". In: *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*. IEEE. 2020, pp. 56–63.

[108] Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. "ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic Levantine Tweets". In: May 2018.

[109] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[110] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[111] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[112] Yoon Kim. *Convolutional Neural Networks for Sentence Classification*. 2014. arXiv: 1408.5882 [cs.CL]. URL: https://arxiv.org/abs/1408.5882.

[113] Abdulaziz M Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. “A combined CNN and LSTM model for Arabic sentiment analysis”. In: *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*. Springer. 2018, pp. 179–191.

[114] Alexis Conneau, Holger Schwenk, Loıc Barrault, and Yann Lecun. “Very deep convolutional networks for text classification”. In: *arXiv preprint arXiv:1606.01781* (2016).

[115] Rie Johnson and Tong Zhang. “Effective use of word order for text categorization with convolutional neural networks”. In: *arXiv preprint arXiv:1412.1058* (2014).

[116] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[117] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

[118] Duyu Tang, Bing Qin, and Ting Liu. “Document modeling with gated recurrent neural network for sentiment classification”. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 1422–1432.

[119] Amal Alharbi, Manal Kalkatawi, and Mounira Taileb. “Arabic sentiment analysis using deep learning and ensemble methods”. In: *Arabian Journal for Science and Engineering* 46 (2021), pp. 8913–8923.

[120] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. “Hierarchical attention networks for document classification”. In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016, pp. 1480–1489.

[121] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359.

[122] Nizar Y Habash. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers, 2010.

[123] Jing Jiang and ChengXiang Zhai. “Instance weighting for domain adaptation in NLP”. In: ACL. 2007.

[124] Giyaseddin Bayrak and Abdul Majeed Issifu. “Domain-adapted BERT-based models for nuanced Arabic dialect identification and tweet sentiment analysis”. In: *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*. 2022, pp. 425–430.

[125] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

[126] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[127] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. “Enriching word vectors with subword information”. In: *Transactions of the association for computational linguistics* 5 (2017), pp. 135–146.

[128] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. “Word translation without parallel data”. In: *International conference on learning representations*. 2018.

[129] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. “Improving language understanding by generative pre-training”. In: (2018).

[130] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. “Unsupervised cross-lingual representation learning at scale”. In: *arXiv preprint arXiv:1911.02116* (2019).

[131] Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. “Pre-Training BERT on Arabic Tweets: Practical Considerations”. In: (2021). arXiv: 2102.10684 [cs.CL].

[132] Wissam Antoun, Fady Baly, and Hazem Hajj. “AraELECTRA: Pre-training text discriminators for Arabic language understanding”. In: *arXiv preprint arXiv:2012.15516* (2020).

[133] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. “On the cross-lingual transferability of monolingual representations”. In: *arXiv preprint arXiv:1910.11856* (2019).

[134] Mashael Al-Duwais, Hend Al-Khalifa, and Abdulmalik Al-Salman. “A Benchmark Evaluation of Multilingual Large Language Models for Arabic Cross-Lingual Named-Entity Recognition”. In: *Electronics* 13.17 (2024). URL: https://www.mdpi.com/2079-9292/13/17/3574.

[135] Abdelrahman Kaseb and Mona Farouk. *SAIDS: A Novel Approach for Sentiment Analysis Informed of Dialect and Sarcasm*. 2023. arXiv: 2301.02521 [cs.CL]. URL: https://arxiv.org/abs/2301.02521.

[136] Walid Magdy brahim Abu Farha Wajdi Zaghouani. “Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic”. In: *The Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics (ACL). 2021, pp. 296–305.

[137] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. “Don't stop pretraining: Adapt language models to domains and tasks”. In: *arXiv preprint arXiv:2004.10964* (2020).

[138] Zhun Deng, Linjun Zhang, Kailas Vodrahalli, Kenji Kawaguchi, and James Y Zou. “Adversarial training helps transfer learning via better representations”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25179–25191.

[139] Yaroslav Ganin and Victor Lempitsky. “Unsupervised domain adaptation by backpropagation”. In: *International conference on machine learning*. PMLR. 2015, pp. 1180–1189.

[140] Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. “Bridging the gap between training and inference for neural machine translation”. In: *arXiv preprint arXiv:1906.02448* (2019).

[141] Nizar Habash, Mona T Diab, and Owen Rambow. “Conventional orthography for dialectal Arabic.” In: *LREC*. 2012, pp. 711–718.

[142] Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. “Overview for the first shared task on language identification in code-switched data”. In: *Proceedings of the first workshop on computational approaches to code switching*. 2014, pp. 62–72.

[143] Mohamed Aly and Amir Atiya. “Labr: A large scale arabic book reviews dataset”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2013, pp. 494–498.

[144] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *arXiv preprint arXiv:1910.01108* (2019).

[145] Alireza Ghorbanali, Mohammad Karim Sohrabi, and Farzin Yaghmaee. “Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks”. In: *Information Processing & Management* 59.3 (2022), p. 102929.

[146] Omar Alharbi. “Classifying Sentiment of Dialectal Arabic Reviews: A Semi-Supervised Approach”. In: *International Arab Journal of Information Technology* 16 (Apr. 2019), pp. 995–1002.

[147] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[148] Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. “The interplay of variant, size, and task type in Arabic pre-trained language models”. In: *arXiv preprint arXiv:2103.06678* (2021).

[149] Ibrahim Abu Farha and Walid Magdy. “From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset”. In: *The 4th Workshop on Open-Source Arabic Corpora and Processing Tools*. European Language Resources Association (ELRA). 2020, pp. 32–39.

[150] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. “Integrating multimodal information in large pretrained transformers”. In: *Proceedings of the conference. Association for computational linguistics. Meeting*. Vol. 2020. 2020, p. 2359.

[151] Yik Tan, Chee Onn Chow, Jeevan Kanesan, Joon Huang Chuah, and YongLiang Lim. “Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning”. In: *Wireless Personal Communications* 129 (Mar. 2023).

[152] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. “Fast wordpiece tokenization”. In: *arXiv preprint arXiv:2012.15524* (2020).

[153] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. "XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond". In: *arXiv preprint arXiv:2104.12250* (2021).

[154] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[155] Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. "An empirical study of pre-trained transformers for Arabic information extraction". In: *arXiv preprint arXiv:2004.14519* (2020).

[156] Ali Safaya. *Arabic-ALBERT*. Version 1.0.0. Aug. 2020. URL: https://doi.org/10.5281/zenodo.4718724.

[157] 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)*. Ed. by Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 502–518. URL: https://aclanthology.org/S17-2088.

[158] Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. "NADI 2021: The Second Nuanced Arabic Dialect Identification Shared Task". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*. 2021, pp. 244–259.

[159] Stamatios-Aggelos Alexandropoulos, Sotiris Kotsiantis, and Michael Vrahatis. "Data preprocessing in predictive data mining". In: *The Knowledge Engineering Review* 34 (Jan. 2019).

[160] Mike Schuster and Kaisuke Nakajima. "Japanese and Korean voice search". In: Mar. 2012, pp. 5149–5152.

[161] Emmanuel D. Bisong. *Google Colaboratory*. https://colab.research.google.com/. Accessed: [date]. 2019.

[162] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019.

[163] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[164] Rasha Obeidat, Amjad Albashayreh, and Lojin Bani Younis. "The Impact of Combining Arabic Sarcasm Detection Datasets On The Performance Of BERT-based Model". In: July 2022.

[165] Obeida Eljundi, Wissam Antoun, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. “hULMonA (حلمنا): The Universal Language Model in Arabic”. In: July 2019.

[166] Mohamed Abdelhakim, Bingquan Liu, and Chengjie Sun. “Ar-PuFi: A short-text dataset to identify the offensive messages towards public figures in the Arabian community”. In: *Expert Systems with Applications* 233 (2023), p. 120888. URL: https://www.sciencedirect.com/science/article/pii/S0957417423013908.