الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

جامعة غرداية
University of Ghardaia
كلية العلوم والتكنولوجيا
Faculty of Science and Technology
قسم الرياضيات والإعلام الآلي
Department of Mathematics and Computer Science
مخبر الرياضيات والعلوم التطبيقية
Mathematics and Applied Sciences Laboratory

Registration n°:
...../...../...../...../.....

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master

Domain: Mathematics and Computer Science
Field: Computer Science
Specialty: Intelligent Systems for Knowledge Extraction

Topic

Corpus Construction for Arabic Question Answering Subjectivity Classification

---

Presented by:

Soumia SOUFFI & Mounia BOUAMEUR

Publicly defended on June 25, 2023

Jury members:

Mr. Slimane Oulad-Naoui	MCB	Univ. Ghardaia	President
Mr. Houssem Eddine Degha	MCB	Univ. Ghardaia	Examiner
Mr. Abdelkader Bouhani	MAA	Univ. Ghardaia	Examiner
Mr. Attia Nehar	MCB	Univ. Z.A. Djelfa	Supervisor
Mr. Slimane Bellaouar	MCA	Univ. Ghardaia	Co-Supervisor

Academic Year: 2022/2023

Acknowledgment

First and foremost, we would like to express our deepest gratitude to Allah, the most merciful and compassionate, for his blessings and guidance throughout our journey.

We are immensely thankful to our supervisors, ***Dr. Nehar Attia*** and ***Dr. Slimane Bellaouar***, for their invaluable guidance, constant encouragement, and scholarly insights. Their expertise and dedication have been instrumental in shaping the direction of our research and refining our ideas. We are truly fortunate to have had the opportunity to work under their mentorship.

We would like to extend our heartfelt gratitude to our family members, for their love, support, and understanding. Their encouragement, and moral support. Their belief in our abilities has been a constant source of strength for us.

We would like to extend our sincere gratitude to the esteemed members of the jury for accepting our work and granting us the opportunity to present it. We deeply appreciate your time, effort, and valuable feedback. we are honored to have received your approval.

We would also like to express our gratitude to the faculty members, administrative staff, and colleagues at Gharadaia University, who have provided a conducive academic environment and enriching opportunities for learning and growth.

Finally, we extend our heartfelt appreciation to all the individuals who, in one way or another, have contributed to the completion of this thesis. Their support, assistance, and encouragement, whether through discussions, feedback, or other means, have played an integral role in the success of our research.

In conclusion, this thesis is the result of collective efforts, and we are deeply indebted to everyone who has been a part of this journey. May Allah bless you all abundantly for your contributions, guidance, and support.

Dedication

I would like to dedicate this thesis to the important people who have been there for me throughout my academic journey.

To my beloved grandfather, **Toubane Ahmed,** and my cherished grandmother, **Rezgui Messouada,** I want to express my deepest gratitude for believing in me and serving me as a constant source of inspiration throughout my academic journey. Your unwavering faith in my abilities has fueled my determination to strive for excellence and achieve success. Your presence in my life has been a blessing beyond measure.

To my incredible **Mother** and **Father**, I am immensely grateful for the sacrifices you have made and the solid support you have provided throughout my academic journey. I cannot thank you enough for your selflessness, guidance, and constant encouragement. Your presence in my life has given me the strength and determination to pursue my dreams.

To my younger siblings, **Mohammed Taher, Saif Eddine Islam,** and **Maria Nourhanne**, thank you for being there to share in both the joys and challenges of my academic journey. I appreciate the unique bond we share as siblings, and I am grateful for your presence in my life.

To my respected **uncles** and my dear **aunties**, I want to express my sincere appreciation for your encouragement and support throughout my academic journey. Your wisdom and mentorship have played a significant role in shaping my path toward success. I am also grateful for the support and friendship of all my cousins, Your presence in my life has made my journey richer and more fulfilling. Thank you all for being an integral part of my academic and personal growth.

To my dear friend and college companion,**Soumia**, Thank you for being by my side throughout our academic journey. Your presence has been a source of strength and support that I am forever grateful for. The memories we have created and the moments we have shared are treasures I hold close to my heart. I am truly blessed to have you as a friend and a sister, and I cherish the bond we have formed.

To my best friends, **Zahra, Fatima**, and **Kawther**, I want to thank you from the bottom of my heart for your unwavering support and belief in me. Your friendship has been a constant source of strength throughout my academic journey.

I am grateful to each and every one of you for your contributions to my success. This thesis is dedicated to all of you.

BOUAMEUR MOUNIA

Dedication

In deep gratitude and with utmost appreciation, I dedicate this thesis to:

To my dearest **Mother,** You have been my pillar of strength throughout this journey. Your love, encouragement, and unwavering belief in my abilities have guided me every step of the way. Your sacrifices and constant support have shaped me into the person I am today. This thesis is dedicated to you, as a testament to your endless devotion and as a token of my deepest gratitude. Thank you for always inspiring me to reach for the stars.

To my loving **Father,** Your wisdom, guidance, and boundless optimism have been the driving force behind my pursuit of knowledge. Your faith in my dreams and your countless hours of discussions have broadened my horizons and ignited my passion. This thesis stands as a tribute to your enduring belief in me and the countless sacrifices you have made to see me succeed. Thank you for being my constant source of inspiration.

To my dear brothers, **Aissa, Abdelbadie, Abdallah** and **Ahmed** my partners in laughter, this thesis is dedicated to all of you as a symbol of our unbreakable bond. Thank you for always cheering me on.

To my dear **Fiancé,** this thesis is dedicated to you, Thank you for always being there for me, for lending a listening ear during moments of frustration, and for supporting me to achieve more.

To my dear friends, **Chiraz, Dalia, Yasmine, Amina,** and **Zahra** , who have stood by me through thick and thin, and filling my life with joy and laughter. Your friendship has been a constant source of inspiration and strength throughout my academic journey. I dedicate this thesis to all of you as a tribute to the beautiful memories we have created together.

To **Mounia** my dear friend and partner of this thesis, I am grateful for the countless hours we spent collaborating, the late-night study sessions, and overcoming challenges together. This thesis is dedicated to you, as a testament to our successful partnership.

To all my family members, I dedicate this thesis to each and every one of you.

To all who have played a part in shaping my journey, whether through a word of encouragement, a listening ear, or a helping hand, you have all contributed to my growth and success. This thesis is dedicated to each and every one of you.

SOUFFI SOUMIA

مـــلـــخـــص

يعتبر تحليل الذاتية والمشاعر موضوعًا مهمًا في مجال معالجة اللغة الطبيعية نظرًا لقدرتها على استخلاص وتصنيف المعلومات الذاتية المعبرة في البيانات النصية. على الرغم من البحوث الواسعة التي أُجريت للغات الرئيسية مثل الإنجليزية، تفتقر اللغة العربية، مع تنوعها اللهجي، إلى الموارد والبحوث الكافية في هذا المجال. يهدف هذا البحث إلى التغلب على نقص الموارد في تحليل الذاتية في اللغة العربية من خلال بناء مجموعة بيانات شاسعة للأسئلة والإجابات باللغة العربية مصممة خصيصًا لتحليل الذاتية. يشمل بناء المجموعة الخطوات التالية: جمع البيانات من خلال استخلاصها من الويب، وتنظيف البيانات لضمان الجودة، تليها عملية التعليق بوضع علامات ذاتية باستخدام نموذجين قمنا بتطويرهما باستخدام تقنية التعديل الدقيق على نموذجين لغويين ضخمين وهما XLM-RoBERTa, AraBERT و تتيح توافر هذه المجموعة تحفيز المزيد من البحث وتعزيز التطورات في معالجة اللغة العربية، وتسهم في تطبيقات متعددة في تحليل المشاعر واستخراج الآراء.

كلمات مفتاحية: تحليل الذاتية، تحليل المشاعر، التعديل الدقيق، AraBERT، XLM-RoBERTa.

Abstract

Subjectivity and sentiment analysis, have gained significant attention in the field of Natural Language Processing (NLP) due to their ability to extract and classify subjective information expressed in textual data. Although, extensive research has been conducted on major languages such as English, Arabic with its dialectal variations lacks sufficient resources and research in this domain. This study aims to overcome the scarcity of resources in Arabic subjectivity analysis by constructing an extensive Arabic Question-Answering (QA) corpus specifically designed for subjectivity analysis. The corpus construction involves the following steps: data collection through web scraping, and data cleaning to ensure quality, followed by the annotation process by affecting subjectivity labels using two models that we developed utilizing the fine-tuning technique with two pre-trained models, XLM-RoBERTa and AraBERT. The availability of this corpus stimulates further research, drives advancements in Arabic NLP, and contributes to various applications in sentiment analysis and opinion mining.

Keywords: Subjectivity analysis, sentiment analysis, fine-tuning, AraBERT, XLM-RoBERTa.

Résumé

L'analyse de la subjectivité et des sentiments a suscité une attention considérable dans le domaine du traitement automatique du langage naturel (TAL) en raison de leur capacité à extraire et classifier les informations subjectives exprimées dans les données textuelles. Bien que des recherches approfondies aient été menées pour les langues principales telles que l'anglais, l'arabe, avec ses variations dialectales, manque de ressources et de recherches suffisantes dans ce domaine. Cette étude vise à surmonter le manque de ressources dans l'analyse de la subjectivité en arabe en construisant un vaste corpus de questions-réponses arabes spécifiquement conçu pour l'analyse de la subjectivité. La construction du corpus implique l'utilisation des étapes suivantes : la collecte de données via le web scraping et le nettoyage des données pour garantir leur qualité, suivi du processus d'annotation en affectant des étiquettes de subjectivité à l'aide de deux modèles qu'on a développés en utilisant la technique de fine-tuning avec deux modèles pré-entraînés, AraBERT et XLM-RoBERTa. La disponibilité de ce corpus stimule de nouvelles recherches, favorise les avancées dans le TAL en arabe et contribue à diverses applications dans l'analyse des sentiments et l'extraction d'opinions..

Mots clés: analyse de subjectivité, analyse de sentiment, fine-tuning, Ara-BERT, XLM-RoBERTa .

Contents

List of Figures

List of Tables

Introduction

In recent years, subjectivity and sentiment analysis (SSA) has gained significant attention in the field of natural language processing (NLP). These areas focus on the understanding of subjective information, such as opinions, emotions, and attitudes expressed in textual data. Subjectivity and sentiment analysis techniques aim to extract and classify subjective information to gain insights into people's sentiments, preferences, and beliefs. With the exponential growth of user-generated content on social media platforms and online forums, accurately analyzing opinions has become crucial for various applications, including market research, opinion mining, and customer feedback analysis.

Subjectivity and sentiment analysis have been extensively studied for major languages such as English. However, Arabic language lacks sufficient annotated data sets and lexicons and has relatively limited research, despite its widespread use across the Arab-speaking world. With its unique linguistic features, Arabic exhibits a high degree of ambiguity due to its rich morphological complexity and dialectal variations making it difficult to develop effective computational models and resources for analyzing subjective content in Arabic text.

The main goal of this study is to overcome the lack of resources in Arabic subjectivity and sentiment analysis by creating an extensive corpus of Arabic question-answer pairs specifically designed for subjectivity. The Arabic QA corpus will be annotated with subjectivity labels using an automated approach. This will involve developing a subjectivity classification model that will be trained on existing labeled data, including publicly available data sets in Arabic. Through the construction of this Arabic QA corpus, this research aims to fill the existing gap in resources and enable researchers to advance the understanding of subjective language use in Arabic and develop more effective models for Arabic NLP applications.

Despite the limited research on subjectivity and sentiment analysis in Arabic, there have been notable advancements in recent years. Existing research on subjectivity classification in Arabic has predominantly relied on traditional, machine learning, and deep learning methods. Traditional approaches typically employ lexicon-based and morphological-based features Abdul-Mageed et al. (2011), Abdul-Mageed et al. (2014), Awwad & Alpkocak (2016). On the other hand, machine learning approaches have demonstrated exceptional results using only three classifiers consistently: Support Vector Machine (SVM), k-Nearest Neighbor (KNN), and Naive Bayes (NB) Duwairi & El-Orfali (2014). However, with the advancements in deep learning techniques, there has been a noticeable shift toward the utilization of Recurrent Neural Networks (RNNs) Alhumoud & Al Wazrah (2022), ensemble methods Alharbi et al. (2021), El Karfi & El Fkihi (2022), and fine-tuning pre-trained language models Alduailej & Alothaim (2022). These newer approaches leverage

the power of deep learning to enhance the accuracy and effectiveness of subjectivity classification in Arabic.

The contribution of this study can be divided into two main aspects, first, building subjectivity classification models, where two pre-trained models were selected and fine-tuned by training them on a collection of datasets that were carefully curated to have a balance of subjectivity examples and augmented to enhance the model's understanding of various subjective expressions. The second aspect is the construction of the Arabic QA corpus which focused on subjectivity involved collecting question-answer pairs from Arabic platforms. The collected data were passed through a cleaning process to remove any noise or irrelevant information that could potentially affect the quality and accuracy of the corpus. Once the cleaned QA pairs were obtained, subjectivity labels were assigned to them using the two developed subjectivity classification models to annotate each QA pair as subjective or objective.

This thesis is divided into four chapters. In Chapter 1, we begin by providing a comprehensive definition of subjectivity and sentiment analysis. We also introduce the concept of large language models and question answers systems (QA). Chapter 2 explores the existing research on sentiment and subjectivity classification, as well as question-answering corpora in both English and Arabic languages. It presents an overview of the latest advancements in these areas, highlighting the techniques, methodologies, and achievements of previous studies. In Chapter 3, we focus on the construction of subjectivity classification models. We delve into the details of the methods, techniques, and tools utilized in building these models. This includes a comprehensive explanation of data preparation, experiments, and evaluation of the models through a discussion of the obtained results. Chapter 4 focuses on the construction of the Arabic Question Answering (AQA) corpus. It provides a detailed overview of the step-by-step process involved in building the corpus, encompassing the collection methods, cleaning process, and data annotation. In the end, we give the conclusions and some perspectives on our work.

Chapter 1

Basic Concepts

1.1 Introduction

In this chapter, we explore essential aspects of NLP and machine learning: subjectivity and sentiment analysis, large language models, and Question-Answering systems. Initially, we discuss the subjective nature of language and how it affects our interpretation and understanding of the text. We delve into the concept of sentiment analysis and how it can be used to identify and classify emotions and opinions expressed in text data. Subsequently, our attention shifts toward the significance of large language models within the domain of NLP. We discuss how these models work, their training process, and their applications in various fields. We also take a closer look at Question-Answering Systems (QAS).

1.2 Subjectivity and Sentiment Analysis

Subjectivity in natural language refers to linguistic features that convey opinions, sentiments, assessments, and hypotheses Abdul-Mageed et al. (2011). It can be stated explicitly, as in reviews or discussion articles that openly utilize subjective language, or implicitly, as in attitude, when the writer of the work adopts a certain perspective on a specific topic and employs its metaphors and words.

A subjective observation can be influenced by several factors, including many different kinds of biases. The first, framing bias, is manifested by subjective terms (adjectives, adverbs, etc.) or expressions associated with a specific point of view, these subjective intensifiers give directional power to a proposition's meaning and so reflect the author's attitude on a certain topic. The second, known as epistemological bias, includes statements that are generally accepted as true or often considered untrue and that are quietly presupposed, indicated, or assumed in the text Recasens et al. (2013). However, in contrast to subjective concepts, objective concepts express some information about the world's reality Liu et al. (2010).

The process of subjectivity classification refers to the task of classifying texts as either objective or subjective,Abdul-Mageed et al. (2011) This classification would be beneficial for many natural language processing (NLP) applications like sentiment analysis (SA) or opinion mining (OM), which is the study of people's attitudes,

thoughts, and feelings regarding individuals, events, or topics Medhat et al. (2014). For sentiment classification, the task consists of analyzing the subjective text and ascertaining its sentiment polarity (e.g., positive, negative, or neutral).

Subjectivity and Sentiment Analysis (SSA) can be formally defined by two stages of classification: (i) The initial stage (i.e., subjectivity classification) is where a given text t from a text set T will be classified by assigning a label s from a set $S = \{subjective, objective\}$. (ii) The second stage (i.e., sentiment classification) is classifying all the t labeled with only s = subjective by assigning polarity labels p from a set of polarities $P = \{positive, negative, neutral\}$ Alotaibi (2016).

1.3 Large Language Models

Large Language Models (LLMs) are fundamental models in Natural Language Processing (NLP) applications, based on deep learning neural networks Transformer architecture that uses the self-attention mechanism to help the model understand the complexity and the semantic connections of the language's vocabulary Kasneci et al. (2023); Vaswani et al. (2017), the table shows some of the most used LLMs.

Table 1.1: Examples of Large language models

Model Name	Parameters Size	Developer	Release year
BLOOM Scao et al. (2022)	176 billion	Hugging Face with Big Science	2022
GPT-3 Brown et al. (2020)	175 billion	OpenAI	2020
BERT Devlin et al. (2018)	340 million	Google	2018
XLM-RoBERTa Conneau et al. (2019)	127 million	Meta	2019

LLMs are usually pre-trained on a large set of unlabeled data to give the model a general understanding of the language and then be further trained (fine-tuned) on a set of labeled data to produce more accurate output prediction for the intended NLP application (See Figure 1.1). Despite the fact that pre-training is far more computationally expensive than fine-tuning, it only has to be done once, moreover, the pre-trained model may be fine-tuned for plenty of language tasks Guu et al. (2020).

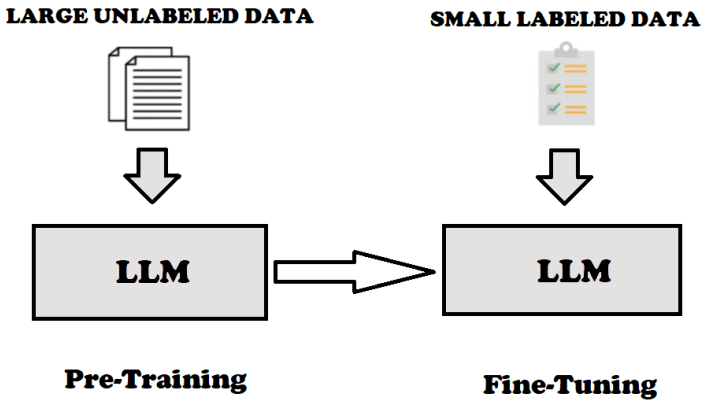

Figure 1.1: Pre-training vs fine-tuning.

Once the model has been pre-trained, it can be trained with task-specific new data to fine-tune it for specific use cases (See Figure 1.2).

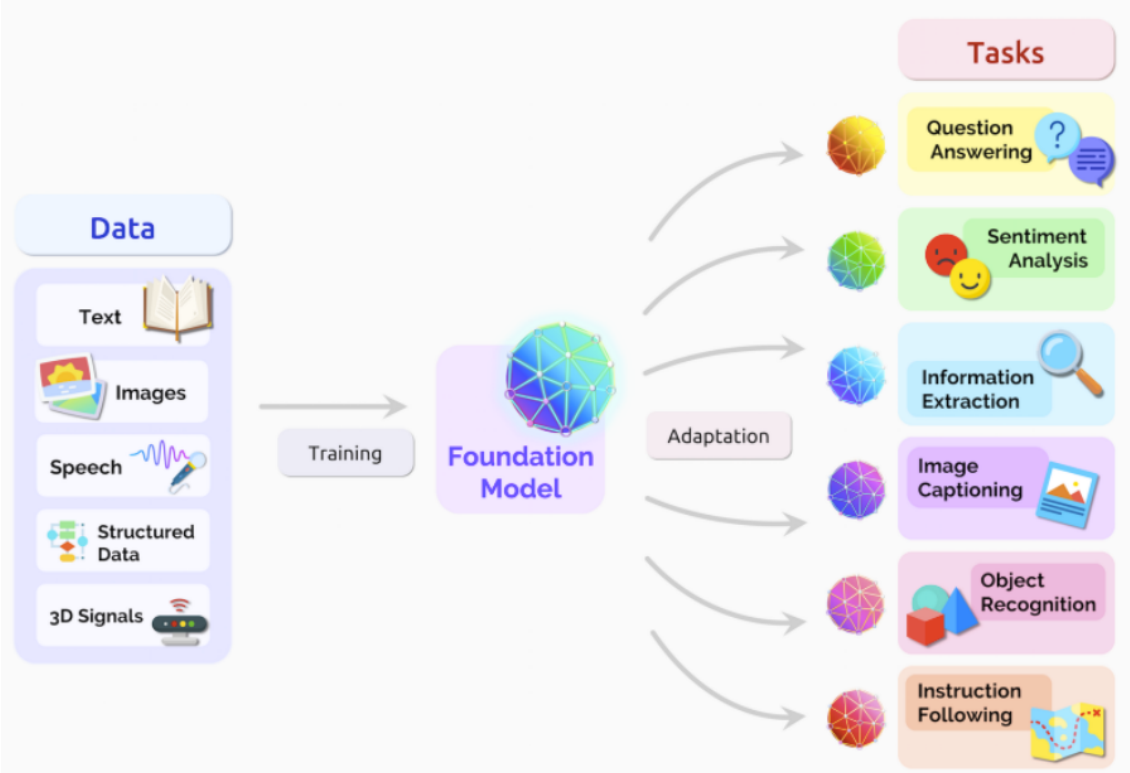

Figure 1.2: Large language models usescases Bommasani et al. (2021).

1.3.1 AraBERT

AraBERT is an Arabic language representation model based on the BERT Devlin et al. (2018) LLM (a stacked Bidirectional Encoder Representations from Transformer), that has been pre-trained particularly for Arabic in order to accomplish the success that BERT achieved in English.
Basically, BERT was trained on 3.3 billion words from English Wikipedia and Book Corpus. However, due to the lack of Arabic resources, AraBERT was trained on only 70 million sentences extracted from two large corpora, the 1.5 billion words Arabic Corpus and the Open Source International Arabic News Corpus (OSIAN). AraBERT architecture is based on the original BERT architecture but specifically designed and trained for the Arabic language. It consists of 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence length, and a total of 110M parameters. They also employ the Next Sentence Prediction (NSP) task that helps the model understand the relationship between two sentences, which can be useful for many language understanding tasks Antoun et al. (2020). The figure 1.3 shows an overview of the BERT architecture:

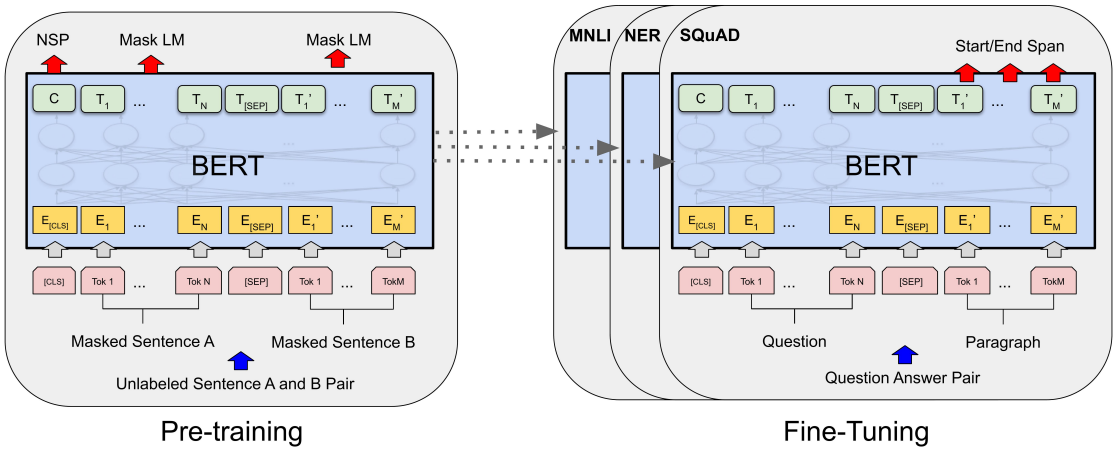

Figure 1.3: Overall BERT's architecture for pre-training and fine-tuning Devlin et al. (2018).

1.3.2 XLM-RoBERTa

XLM-RoBERTa is a multilingual model based on RoBERTa (a transformer model pre-trained on a large corpus in a self-supervised fashion), and it was pre-trained with the Masked Language Modeling (MLM) objective which is a form of denoising autoencoding, where a model is trained to restore a corrupted input sequenceWettig et al. (2022), On 2.5 TB of filtered CommonCrawl data containing 100 languages including Arabic. XLM-R uses the same MLM objective as the XLM model with only one change which is removing the language embeddings which allows the model to better deal with code-switching. This way, the model can learn useful representations of languages to produce helpful features for specific tasks Conneau et al. (2019).

An advantage of XLM-RoBERTa's multilingual pre-training is that it can be fine-tuned for specific downstream tasks, such as text classification. A simplified

example of the architecture specific to text classification can be observed in the figure 1.4

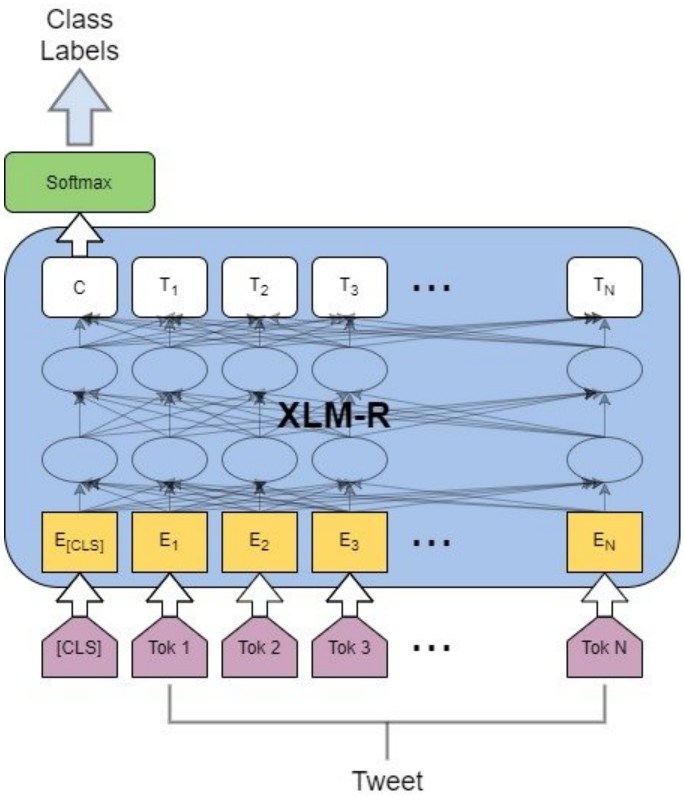

Figure 1.4: Example of text classification architecture with XLM-R Ranasinghe & Zampieri (2020).

1.4 (Arabic) Question Answering Systems

Question-answering systems are systems that combine various fields from NLP, information retrieval (IR), and information etraction (IE) to provide accurate and relevant answers to questions posed. The first step in a QA system involves question classification, where the system determines the type of question being asked to better understand the user's intent. Information retrieval then comes into play, where the system searches through relevant text sources, such as online articles or documents, to retrieve the most suitable information for answering the question. Finally, answer extraction techniques are employed to extract the relevant information from the retrieved sources and present it to the user as a concise and accurate answer, Allam & Haggag (2012).

Arabic question-answering systems have garnered substantial attention due to the widespread usage of Arabic as a natural language and the significant increase in content on the Internet and the demand for information that regular information retrieval approaches cannot provide[Shaheen & Ezzeldin (2014).

The development of robust Arabic question-answering systems has the potential to greatly improve access to information and facilitate knowledge dissemination for Arabic speakers worldwide.

1.5 Conclusion

The fundamental concepts covered in this chapter are essential building blocks for a strong understanding and solving NLP and ML tasks. In addition to defining the basics terms, we have also explored commonly used tools and techniques such as XLM-R and Ara-BERt that represent a crucial part of our work.

Chapter 2

State Of The Art

2.1 Introduction

This chapter includes a compilation of relevant works to our topic, beginning with papers on subjectivity and sentiment classification in both English and Arabic, categorized by traditional, machine learning, and deep learning methodologies. We then review some English and Arabic Question Answering Corpora.

2.2 Subjectivity and Sentiment Classification (SSC)

2.2.1 SSC Systems for English

Researchers have tackled the challenge of analyzing subjective content and sentiment in English texts using various methodologies, and approaches stated as follows.

Traditional approaches

By conducting an experimental investigation on a corpus of product evaluations, the authors of Wu et al. (2009) define the phrase dependency parsing methodology and demonstrate its efficacy in the task of opinion mining. The results suggest that this strategy improved task performance.

Later in Kouloumpis et al. (2011), the authors investigate whether part-of-speech (POS) tags and sentiment lexicons are beneficial for sentiment analysis on Twitter by conducting a study on three independent corpora of Twitter messages. They use the hashtagged dataset (HASH), the emoticon dataset (EMOT), and a manually annotated dataset from iSieve Corporation. In addition, they incorporate the aspect of POS tagging during the data preprocessing step. According to the study's findings, characteristics from an existing sentiment lexicon are more useful for sentiment analysis on Twitter than POS features.

Machine learning approaches

The challenges of the sentiment analysis task at the document level are discussed in Pang et al. (2002), where they attempt to develop an effective SA model using movie reviews to train multiple classifiers such as Naive Bayes, Maximum entropy, and SVM. They compare them to human baselines. They discovered that machine learning techniques outperform humans in SA tasks.

In the domain of sentence subjectivity classification, B. Wang et al. (2008) investigate a semi-supervised learning method, self-training, by adopting decision tree models such as C4.5, C4.4, and Naive Bayes tree (NBTree) as underlying classifiers and an adapted Value Difference Metric (VDM) as the selection metric in self-training.
To evaluate the effectiveness of this semi-supervised learning strategy based on various underlying classifiers leveraging multiple selection metrics under various conditions, a number of tests on the MPQA corpus were developed. According to the obtained results, self-training using the underlying classifiers VDM and NBTree performs better than self-training using other different combinations of underlying classifiers and selection metrics. They demonstrate that the investigated strategy may perform in parallel with supervised learning methods for sentence subjective classification.

Deep Learning approach

W. Wang et al. (2017) propose a deep learning model called couples multi-large attention (CMLA) for the task of aspect and opinion term co-extraction. This model is based on an end-to-end network instead of the syntactic parsing of each sentence, which is an effective way to complete this task but costs a lot of effort. The proposed model was evaluated on 3 datasets which proved the efficiency of the CMLA model in this task.

Ghosal et al. (2018) provide an RNN model based on multi-modal attention that employs contextual information for SA classification. The developed model was verified on two different multi-modal sentiment datasets: the CMU-MOSI corpus for sentiment intensity and the CMU-MOSEI corpus for emotional intensity. The results demonstrate the efficacy of the suggested model on corpora.

A completely new architecture (VDCNN) for natural language processing (NLP) is built by the authors of Conneau et al. (2016). It functions directly at the character level and employs a deep stack of local operations including convolutions and pooling operations. This architecture has been tested on eight publicly accessible large-scale data sets and the results reveal that the performance increases consistently when the depth is increased up to 29 convolutional layers. On top of that, the results reflect improvements above a number of text classification tasks.

Chen et al. (2016) present a hierarchical long-short-term memory (LSTM) model that combines user and product information as attention over several semantic levels of a document to address the challenge of document-level sentiment classification, which is an important topic in sentiment analysis. The performance of this model has been evaluated using tests on several actual data sets and the results indicate that the proposed model consistently and significantly outperforms

previous models.

The authors of Giannakopoulos et al. (2017) propose a B-LSTM and CRF classifier that was trained using labeled datasets that were automatically classified. The classifier may be utilized to extract features and identify aspect terms in texts for both supervised and unsupervised aspect term extraction (ATE). On the dataset of the SemEval-2014 Aspect Based Sentiment Analysis (ABSA) contest for the laptop and restaurant domains, they validated their classifier with supervised ATE and top-ranking performance.

2.2.2 SSC Systems for Arabic

Similar to the exploration of subjectivity classification in the English language, we classify the studies into the following categories.

Traditional approaches

In Abdul-Mageed & Diab (2014), the authors created SANA a large-scale lexicon for Arabic and Arabic dialects that was annotated by involving two approaches: manual and automatic.

In the manual approach, they included human annotation to label two distinct genre lexicons: SIFAAT was collected from the first four parts of PATB, and HUDA was obtained from the Egyptian chat room of Yahoo Maktoob!.

In the automatic approach, they adopted two techniques: a static technique using PMI (point view mutual information) and another technique using google translate API to translate existing English labeled lexicons: WordNet, SentiWordNet (SWN3), Youtube Lexicon (YT), and Affect Control Theory Lexicon (ACT) to Arabic. SANA is developed using several genres (Online news wire, chat turns, Twitter tweets, YouTube comments) and dialects and languages, which is advantageous for Arabic SSA.

Awwad & Alpkocak (2016) provide an SA research on four different Arabic lexicons: HarvardA: translation of the Harvard IV-4 Dictionary, two versions of translated MPQA subjectivity lexicon developed by Pittsburgh University and HRMA: a mixture of HarvardA and MPQAII.

They attempt to measure the efficiency of these lexicons on three datasets. LABR: Large Arabic Books Reviews; TA: Twitter dataset for SA; and PatientJO: a health sector dataset built by the authors of this paper. Where they constructed an Arabic lexicon-based analyzer (ALBA) for sentiment classification at both the document level and sentence level.

The experiments have revealed that the best lexicon for the LABR dataset is HRMA, whereas the best for the PatientJO dataset is HarvardA. The results also prove that the lexicon-based technique at document-level and sentence-level generates equivalent results.

In a separate contribution, Abdul-Mageed & Diab (2012) introduced an annotated multi-genre sentence-level corpus for SSA in MSA called AWATIF, built from

different resources including PATB, Wikipedia Talk Pages, and Web forums.

They adopted two labeling methods: simple (SIMP) and linguistic standards (LG) with the assistance of trained annotators (GH) and Amazon Mechanical Turk (AMT) in three distinct situations: LG-GH, SIMP-GH, and SIMP-AMT.

Through the annotation process, they attempted to elaborate on the importance of LG rules on the SSA tasks and assess the effectiveness of both human annotation and crowd-sourcing techniques. As a result, the incorporation of LG guidelines demonstrates significant improvements in annotation quality.

The AWATIF corpus is expected to fill the research gap in the subjectivity and sentiment analysis systems for Arabic.

Abdul-Mageed et al. (2011) developed a sentence-level SSA system for MSA by applying language-independent features and Arabic-specific morphology-based features through studies that they performed on three preprocessed manually annotated news wire genre texts from PATB (Penn Arabic Tree Bank).

They observed that the inclusion of morphology-based features enhances the system's performance.

Later, Abdul-Mageed et al. (2014) built another sentence-level SSA system called SAMAR, but this time for Arabic social media texts, taking into account four distinct genres that encompass both modern standard Arabic and dialect: chat, Twitter, web forums, and Wikipedia talk pages.

In this study, the authors intend to analyze four research aspects: the effects of morphology on Arabic SSA, the usage of standard features, managing the existence of dialects, and finally retrieving the best social media-specific features.

As general results, the employing of both POS sets (ERTS and RTS) optimizes the performance of the SAMAR system, while using standard features only improves subjectivity classification.

Abbasi et al. (2008) Adopted the approach of the language-independent feature for sentiment classification by constructing a set of stylistic and syntactic features. They utilize the root extraction algorithm for the feature extraction and the entropy-weighted genetic algorithm (EWGA) incorporated with the information gain (IG) for the feature selection part.

The Arabic study was conducted on Middle Eastern web forums, where the experiments performed better when using the combined features together than when using just one of them.

Machine learning approach

Al-Rubaiee et al. (2016) aim to provide an Arabic sentiment analysis system for Mubasher products (a Saudi Arabian stock analysis software in Gulf region) through the Twitter platform by selecting MSA and dialect tweets to classify them according to their polarity. After the preprocessing phase, numerous feature selection techniques such as TF-IDF and BTO are used with both Naïve Bayes and SVM classifiers.

The results demonstrate that SVM is superior to Naïve Bayes. Also the authors underline the importance of the preprocessing techniques in SA classification.

The authors of Rushdi-Saleh et al. (2011) present an Arabic opinion corpus OCA for the task of sentiment analysis built from Arabic movie reviews collected from relevant web pages about films. The corpus contains 500 reviews balanced in terms of polarity (250 positive and 250 negative). They also created its English version EVOCA using automatic machine translation to compare the results of different machine learning algorithms and determine the impact of the applied preprocessing techniques on the corpora.

The experiments demonstrate that SVM performs better in both corpora; stemming was not advantageous. Moreover, the EVOCA corpus shows comparable results to the English SA studies.

Duwairi & El-Orfali (2014) examine the performance of three classifiers (SVM, Naive Bayes, and K-nearest neighbor). To analyze the sentiment of Arabic reviews, they employ various preprocessing techniques such as stemming, word n-gram, character n-gram, etc. to two separate datasets of reviews. The first was a brand-new dataset created by the authors of this research called the politics dataset that was gathered from the Aljazeera website, and the second was a publicly accessible dataset called the movie dataset.

The achieved result demonstrates how the preprocessing techniques and dataset influence the performance of the classifier.

Mountassir et al. (2012) evaluate the performance of the SVM, Naive Bayes, and K-nearest neighbor classifiers while exploring several preprocessing techniques, including term frequency thresholding, term weighting, stemming type and n-gram terms. The experiments were conducted on two Arabic corpora OCA (Opinion Corpus for Arabic) a corpus made up of movie reviews gathered from various Arabic blog sites and web pages and labeled automatically And ACOM (Arabic Corpus for Opinion Mining) corpus that was collected from Aljazeera's site and manually labeled, It consists of two datasets: DS1 movie-review domain dataset and DS2 sport-specific dataset.

Based on a comparison of the results obtained on the two corpora, for OCA, The three classifiers have proven to be highly effective. However, for ACOM, NB and SVM were equally effective, while k-NN was less so.

Deep Learning approach

DeepASA, an Arabic sentiment analysis model, was developed in Alharbi et al. (2021) as a combination of deep learning, ensemble methods, and multiple word embedding techniques including word2vec, doc2vec, and FastText.

The experiments were conducted on six different datasets (LABR, HARD, Restaurant reviews, Product reviews, Ar-Twitter, ASTD) to prove the effective performance of DeepASA according to the authors' comparison to other SA works.

Alhumoud & Al Wazrah (2022) present a review of Arabic sentiment analysis systems using RNNs, where they discussed in detail the structure of RNNs architectures like LSTM and GRU and how to improve the prediction results. In addition,

they attempted to demonstrate that the use of deep learning techniques performs better compared to the models built using machine learning.

El Karfi & El Fkihi (2022) developed an ensemble model for Arabic sentiment analysis that combines the majority voting model with the AraBERT and CAMeL-BERT transformer language-based models. To explore the efficacy of this proposed model, the authors train all four models on three datasets: Twitter datasets written in MSA and Jordanian dialect, Gold-Standard Twitter dataset, and ASTD (Arabic sentiment tweets dataset).

They test the proposed model on their balanced manually annotated modern standard Arabic book reviews dataset obtained from the Goodreads website, as well as the three previous datasets.

The findings of this study reveal that the suggested model outperforms the two transformers language-based models and majority voting model, in terms of F1 score and accuracy on the balanced modern standard Arabic books reviews dataset and the balanced ASTD dataset. However, the performance of other models is different from one model to another on the three datasets.

Recently, AraXLNet, an XLNet-based model for Arabic was developed in two parts (Alduailej & Alothaim (2022)). The first stage is to pre-train the state-of-the-art language model XLNet on a huge Arabic corpus that does not require annotation and is composed of many publically accessible datasets: OpenSubtitles, HARD, LABR, BRAD, AraSenTi, SemEval, and AJGT. The second phase is to finetune the pre-trained language model XLNet using annotated Twitter Arabic dataset for sentiment analysis (ASTD).

The results of this experiment reveal that AraXLNet overcame the AraBERT model and SVM in Arabic sentiment analysis tasks utilizing benchmark datasets.

2.3 Question Answering Corpora

QA corpora are essential for advancing research in natural language processing (NLP) and developing effective QA systems. These corpora provide valuable resources for researchers and practitioners working on QA tasks in both Arabic and English languages. In this section, we highlight a few QA corpora available in Arabic and English.

2.3.1 English QA Corpora

TrainQA (Tomás et al. (2009)) a comprehensive English training corpus designed specifically for corpus-based question-answering systems. It offers a rich collection of question-answer pairs, providing essential components such as questions, question types, exact answers, sentences, paragraphs, and document context where the answers can be found. Additionally, the dataset includes labels indicating whether the extracted answer is correct (positive) or incorrect (negative).

It consists of 71,982 samples carefully constructed through a semi-automatic process. Within this corpus, there are 7,598 positive samples, indicating correct

answers, and 64,384 negative samples, representing incorrect answers These samples were meticulously collected to match the requirements of real-world QA systems. The dataset incorporates questions from the TREC competitions held between 2002 and 2005.

TrainQA represents a valuable resource that empowers researchers and practitioners in the field of corpus-based question answering systems.

The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. (2016)) is a large English QA dataset consisting of a set of questions posed by crowd workers on a collection of Wikipedia articles. Each question's response is a part of the corresponding reading passage.

There are two main versions of SQuAD which are: SQuAD1.1 contains more than $100k$ QA pairs on 500 articles and SQuAD2.0 (Rajpurkar et al. (2018)) which is the latest version builds upon SQuAD1.1 by incorporating an additional 50,000 unanswerable questions. This augmentation aims to enhance the training of question-answering systems, enabling them to learn to reject answering when faced with questions that lack suitable responses.

2.3.2 Arabic QA Corpora

DAWQAS (Ismail & Homsi (2018)) is an Arabic dataset of Why (in Arabic: لماذا) question-answering system built to address the lack of non-publicly available datasets for Arabic. It consists of 3205 why question-answer pairs that were taken from six different publicly accessible Arabic websites and survive through the five stages of data construction (whyQA data cleaning, preprocessing, WhyQAs re-categorization, calculation of rhetorical relations probabilities, annotation, and determination of the exact answer's location for each sentence in why-answers).

The comparison between DAWQAS and the other existing Arabic WhyQA datasets shows that DAWQAS is larger than other datasets since it includes eight categories and has a long why-answer.

Another Arabic work in the same context is TALAA-AFAQ (Aouichat & Guessoum (2017)), a corpus of Arabic Factoid Question Answers that was created for the need of training the Arabic QA systems (AQAS).

The corpus construction involves five steps by extracting syntactic and semantic properties and also defining a collection of the answer patterns to each question. It consists of 2002 QA pairs with 618 pairs having associated answer patterns. It is categorized into four primary classes and 34 finer ones. All answer patterns and features have undergone validation by Arabic experts.

This corpus represents the first of its kind for Arabic Factoid Question Answers, tailored to support the development of Arabic QASs (AQAS).

2.4 Conclusion

This chapter has provided a comprehensive compilation of pertinent research in our subject area. We began by exploring papers that delve into subjectivity and sentiment classification, encompassing both English and Arabic languages. These papers were categorized based on the methodologies employed, including traditional, machine learning, and deep learning approaches. Additionally, we delved into the realm of English and Arabic Question Answering Corpora, shedding light on the available resources in these languages.

Chapter 3

Semantic-based Models for Arabic Subjectivity Classification

3.1 Introduction

This chapter focus on our first contribution. We describe the building of a robust and effective tool for the subjectivity classification of Arabic text. To achieve this, we conduct several experiments using different techniques and approaches. We start by explaining the methods, the data set preparation process, and describing the development environment used in our experiments. We conclude by revealing the results, analyzing findings, and discussing their implications for building an effective tool for Arabic subjectivity classification.

3.2 Methods

In this section, we provide an explanation of the methodology employed in the construction of the Arabic subjectivity classification model. To achieve this objective, we adopt two methods.

1. **Method A: Fine-tuning Large Language Models**

   It consists of fine-tuning two LLMs, namely: XLM-RoBERTa [1] and AraBERT [2] on the downstream task of Arabic text subjectivity classification. These LLMs were previously mentioned in Section 1.3. They are trained for this by using a collection of data sets that are detailed in Section 3.3.

   Following the process of fine-tuning, we assess each model's performance to ensure they can perform accurately.

2. **Method B: Parallel Approach**

   The second method is a parallel one where the Arabic text is translated into English first, and then an existing multi-lingual XLM-R model called "fact-

[1] `https://huggingface.co/xlm-roberta-base`

[2] `https://huggingface.co/aubmindlab/bert-base-arabertv2`

or-opinion-xlmr-el" [3] is used. This model has been trained on a mixture of English and Greek 9000 sentences annotated as subjective or objective, and it has been specifically tuned for the task of subjectivity classification in the English language.

We evaluate the model's performance to determine its effectiveness on the Arabic-translated text and to investigate how the translation process may affect the prediction's quality.

3.3 Dataset Preparation

To build our Arabic subjectivity classification models, a large amount of labeled data is required for training and testing. These data should be annotated carefully as either subjective or objective to be able to train and test our models. We compile a large corpus of texts with annotations leveraging already-existing data sets. The process of creating this dataset consists of the following steps:

3.3.1 Data Collection

Apart from the ASTD (Nabil et al., 2015) dataset, we are not aware of any other dataset in which texts are annotated as subjective or objective. All datasets encountered are dedicated to sentiment analysis and classification, in which subjective text is annotated as either positive, neutral, or negative polarity. To overcome this, we consider any text having one of these labels as a subjective one.

For the objective data, we opted to leverage the SANAD dataset (Einea et al., 2019), which is a single labeled large collection of Arabic news articles categorized into one of the following classes: Culture, Finance, Medical, Politics, Religion, Sports and Technology. Articles from the Medical, Sports, and Technology sections are typically regarded as providing objective viewpoints on facts relating to medical science, sports activities and contests, and technology developments, respectively.

In the following paragraphs, we describe each dataset used for building our training dataset.

ASTD (Arabic Sentiment Tweets Dataset) (Nabil et al., 2015) is a dataset for Arabic social sentiment analysis sourced from Twitter. It comprises approximately 10,000 tweets that have been classified into categories including objective, subjective positive, subjective negative, and subjective mixed.

LABR (Large Arabic Books Reviews) (Aly & Atiya, 2013) is an Arabic large sentiment analysis dataset, consisting of over 63,000 book reviews, each rated on a scale of 1 to 5 where ratings of 4 or 5 were considered as positive, ratings of 1 or 2 were considered as negative, and a rating of 3 considered neutral.

HARD (Hotel Arabic-Reviews Data set) (Elnagar et al., 2018) is an Arabic data set, comprises 490,587 hotel reviews collected from the Booking.com website, where the reviews are expressed in Modern Standard Arabic as well as dialectal

[3] `https://huggingface.co/lighteternal/fact-or-opinion-xlmr-el`

Arabic. We considered the balanced version of HARD which consists of 94052 reviews, each review is rated on a scale of 1 to 5 stars divided into positive with ratings of 4 and 5 and negative with ratings of 1 and 2, however, the neutral reviews with a rating of 3 have been removed from this version of the data set.

SANAD (Single-labeled Arabic News Articles Dataset) (Einea et al., 2019) is a large Arabic data set of textual data consisting of 194,797 articles combined from three datasets that were extracted from three news sources, which are AlKhaleej, Akhbarona, and AlArabiya. Articles fall into one of seven categories: Medical, Finance, Culture, Politics, Religion, Sports and Technology.

3.3.2 Data Balancing and Augmentation

The second step in our process is to address class unbalancing problem within the ASTD dataset. The ASTD contains $6,458$ tweets labeled as objective and $3,218$ tweets labeled as subjective. To achieve this goal, we used oversampling techniques Mohammed et al. (2020) where instances from the minority class (subjective) were re-sampled (duplicating instances) to establish a balanced distribution of data between subjective and objective classes.

Balancing the ASTD dataset became a crucial step due to the weak results obtained when fine-tuning the XLM-R model on an unbalanced dataset and evaluating its performance (67% accuracy found). The presence of significant class imbalances within the data posed a challenge to the model's ability to learn and generalize effectively. By balancing the dataset, we aimed to address this issue and create a more equitable distribution of samples across all classes.

To further enhance the dataset, we have incorporated three additional datasets: SANAD, LABR, and HARD. This decision was driven by the motivation to create a larger and more diverse dataset, building upon the encouraging results obtained from fine-tuning the XLM-R model on the oversampled ASTD dataset (results showed in the table 3.4). These positive outcomes served as a strong motivation to aspire for more. By including these datasets, we aimed to expand the range of subjective and objective texts available for training the model.

From the SANAD dataset, we extracted $32,500$ news articles that are considered objective texts. This addition was essential to introduce a substantial number of unbiased samples, allowing the model to learn the distinguishing features of objective language.

In order to maintain the balance of the augmented dataset, we equally selected $32,500$ subjective reviews equally picket from both the LABR and HARD datasets. These datasets were carefully chosen for their relevance and suitability to the subjective language.

By incorporating these additional datasets, we aimed to achieve two main objectives. Firstly, we sought to increase the overall size of the over-sampled version of ASTD, providing a more comprehensive and diverse training set for the model. Secondly, we aimed to maintain the balance between subjective and objective instances within the augmented dataset, ensuring that the model receives adequate exposure to both types of texts.

By expanding the dataset through the inclusion of SANAD, LABR, and HARD datasets, we aimed to further improve the model's performance, generalization capability, and robustness. This augmentation process was a strategic step towards creating a more comprehensive and effective model for subjectivity classification.

Figure 3.1 provides a visual representation of the data balancing and augmentation process for easy reference.

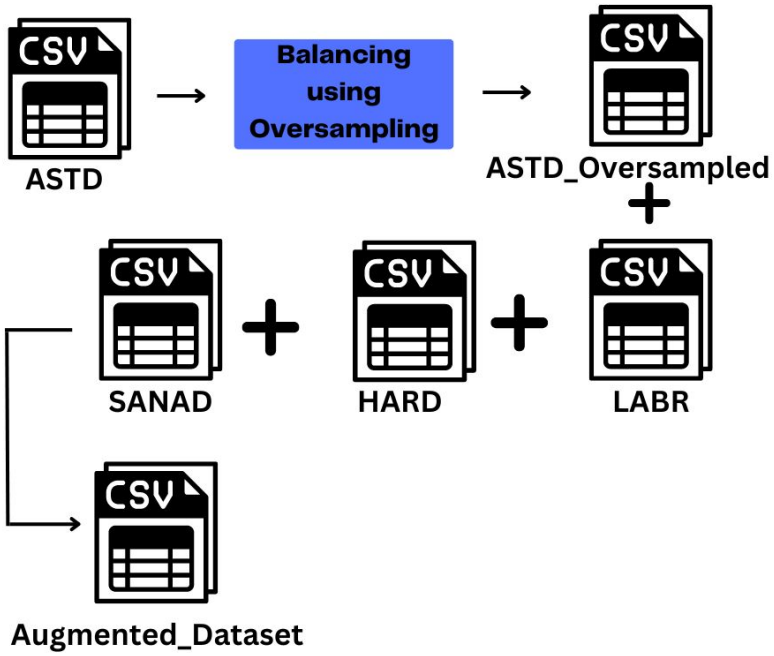

Figure 3.1: Data Balancing and Augmentation process.

3.3.3 Augmented Data Format

The augmented data file is formatted in a commonly used and easily accessible CSV (Comma-Separated Values) format. It is organized in rows and columns, where each row corresponds to a unique text instance, and each column represents a specific attribute as follows:

Text : holds the actual text data.

Class : indicates the subjectivity class for each instance. The subjectivity classes are categorized as pos (positive), neg (negative), neutral, and obj (objective).

Domain : specifies the domain of the text, such as "tweets," "book reviews," "hotel reviews," or "Sports"...etc.

Label : represents subjectivity numerically, with a value of 1 denoting subjectivity and a value of 0 representing objectivity.

In Table 3.1, an instance example of the augmented data is organized and presented. The table showcases the different columns and their corresponding values, providing a clear representation of the augmented dataset.

Table 3.1: Examples of the augmented data.

Text	Class	Domain	Label	Dataset
أهنئ الدكتور أحمد جمال الدين، القيادي بحزب مصر، بمناسبة صدور أولى روايته	POS	Tweets	1	ASTD
فعلا نصائح مميزة ومفيدة جدا احسن حاجة في الأحلام أنها بتكتب عن تجربة	POS	Books reviews	1	LABR
"فندق فاشل". انا حجزت ووصلت فالموعد ولم أجد غرف. ما يصلح فاشل النظام	NEG	Hotel reviews	1	HARD
أعلنت شركة جوتن، إحدى أبرز الشركات العالمية في مجال إنتاج وتوريد الدهان والطلاء وبودرة الطلاء، عن إطلاقها النسخة العربية من موقعها الإلكتروني	OBJ	Technology	0	SANAD

3.3.4 Dataset Splitting

Typically, the collected dataset is split into a training set and a test set. The training set is used to train the model, while the test set is used to evaluate its performance. Our training data represents 80% of the entire data set which compromises 80% of each data set. The remaining 20% of each data set was set aside for the test set by maintaining class balancing and ensuring a random split and representative distribution of the data, The exact number of instances in each set is provided in Table 3.2.

Table 3.2: The train and test statistics.

	Train data	Test data
ASTD	10332	2584
LABR	13000	3250
HARD	13000	3250
SANAD	26000	6500
Total	62332	15584

3.4 Development Environment

We provide insights into the resource and experimental environment in which the experiments are conducted.

Google Colab

Google Colaboratory, commonly referred to as Google Colab, is a cloud-based platform developed by Google, offering users the ability to write, execute, and collaborate on code within an environment tailored for machine learning, data analysis, and educational purposes.
Colab serves as a hosted Jupyter notebook service, eliminating the need for any initial setup, and granting users free access to computing resources, including a GPU runtime complemented by an Intel Xeon 2.20 GHz CPU, 12.68 GB of RAM, and 78 GB of disk space (*Google Colab*, 2023).

Python

Python is a high-level, object-oriented programming language with dynamic semantics that is interpreted. It boasts a rich set of built-in data structures and supports dynamic typing and binding. These features make Python highly desirable for Rapid Application Development, as well as for scripting or integrating existing components.
The language's simplicity and readable syntax contribute to reduced program maintenance costs. Python promotes modularity and code reuse through its support for modules and packages. Additionally, the Python interpreter and comprehensive standard library are freely available in source or binary form for all major platforms, allowing for unrestricted distribution *Python official website documentation. What is Python? Executive Summary* (2023).

Python Libraries

Transformers: is a dedicated library that focuses on supporting Transformer-based architectures and simplifying the distribution of pre-trained models. The core component of the library is an implementation of the Transformer, designed to cater to both research and production needs. The primary aim is to provide robust implementations of popular model variations that are easy to understand, extend, and deploy.

Transformers is an ongoing project diligently maintained by a team of engineers and researchers at Hugging Face. It also benefits from the contributions and active involvement of a thriving community of over 400 external contributors.

The library is released under the Apache 2.0 license and is freely available on GitHub. Detailed documentation and tutorials can be accessed on Hugging Face's website, providing comprehensive resources for utilizing the Transformers library effectively. Wolf et al. (2019).

Pandas: is a Python software library that specializes in data manipulation and analysis. It provides a wide range of data structures and functions specifically designed for manipulating numerical tables and time series data. Pandas is an open-source library, licensed under BSD, which enables users to import data from various file formats including comma-separated values (CSV), JSON, Parquet, SQL database tables or queries, and Microsoft Excel *pandas* (2023).

NumPy : is a fundamental Python package extensively used in scientific computing. It serves as a powerful library that introduces a multidimensional array object, as well as various derived objects like masked arrays and matrice, and an assortment of routines for fast operations on arrays, including mathematical computations, logical operations, shape manipulation, sorting, selection, input/output handling, discrete Fourier transforms, basic linear algebra, statistical operations, random simulation and much more *NumPy documentation* (2023).

PyTorch: is a comprehensive package that encompasses data structures designed specifically for multi-dimensional tensors. It not only defines mathematical operations tailored for these tensors but also offers a wide range of utilities for efficient serialization of tensors and various other types. Moreover, TORCH includes a CUDA counterpart, allowing you to leverage the computational power of NVIDIA GPUs with a compute capability of 3.0 or higher, enabling accelerated tensor computations on these GPUs *TORCH* (2023).

Scikit-learn: commonly referred to as sklearn, is a Python machine learning library available as free software. It offers a comprehensive collection of algorithms for classification, regression, and clustering tasks. These algorithms encompass a wide range of techniques such as support-vector machines, random forests, gradient boosting, k-means, and DBSCAN. Sklearn is specifically designed to seamlessly integrate with other essential Python libraries for numerical and scientific computing, namely: NumPy and Pandas Pedregosa et al. (2011).

3.5 Experiments

Throughout this section, we present the experiments carried out on our dataset following the methodology mentioned earlier in Section 3.2 to build our subjectivity classification model.

Experiment A: Fine-tuning Pre-trained Models

In this experiment, we aim to fine-tune the pre-trained language models, previously mentioned in Chapter 1, on the task of subjectivity classification and evaluate their performance. We use the Hugging Face Transformers (Section 3.4) library to fine-tune the XLM-R and AraBERT models on the Arabic subjectivity classification task using the training set of our augmented dataset.
Fine-tuning the models was done on a GPU platform, using a Mini-Batch Gradient Descent (MGD) with a mini-batch size set to 16 and Adam optimizer with a learning rate of 15e-6. The epoch values used for fine-tuning range from 1 to 5.

The models were evaluated using a testing set from the same data set, and the accuracy, precision, recall, and F1-score were computed.

Initially, as a first attempt to have a subjectivity classification model, we performed a fine-tuning of XLM-R on the over-sampled version of the ASTD dataset, which has been balanced to address the class unbalancing issue. Additionally, a visual representation of the fine-tuning process of the over-sampled ASTD is shown in Figure 3.2. The obtained F1 performance of the resulting model (referred to hereafter as AraSubjXLM-R_1) was about 83%. The other calculated metrics can be found in Table3.4.

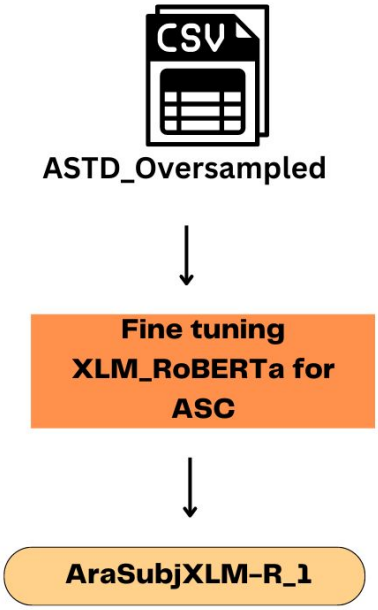

Figure 3.2: Fine tuning XLM_RoBERTa using Oversampled ASTD.

As there are more rooms to enhance this result, we decided to enlarge the over-sampled version of the ASTD dataset by adding three additional datasets: LABR, HARD, and SANAD, which were also pre-processed and labeled for subjectivity classification.

Then, to determine whether the additional data had a beneficial effect on the model's accuracy, a new attempt of fine-tuning of XLM-R on the augmented dataset is launched, Figure 3.3 shows an overall description of this process. The resulting model (referred to hereafter as AraSubjXLM-R_2) achieved an F1 performance of approximately 96%. More details of the obtained results are shown in Table3.5.

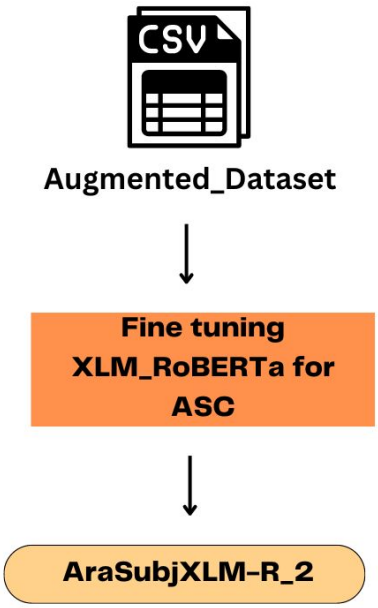

Figure 3.3: Fine tuning XLM_RoBERTa using the augmented dataset.

Another significant part of our experiment (A) involved the fine-tuning of the AraBERT model.

After completing the initial fine-tuning process of XLM-R, we proceeded to apply the same methodology to fine-tune the AraBERT on the augmented version of our dataset (referred to hereafter as AraSubjBERT), Figure 3.4 represents a visual overview of the fine tuning process . A slight sight of results found is F1 score of 97%. The model's performance is detailed in the table 3.6.

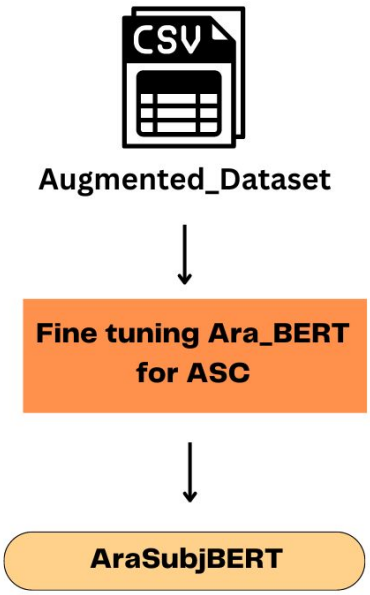

Figure 3.4: Fine tuning Ara_BERT using the augmented dataset.

The code snippet in Figure 3.5 represents a typical training loop for fine-tuning process. It follows the standard steps of training a model, including forward and backward passes, updating the model parameters, and evaluating the performance on both the training and test datasets.

```
from tqdm import tqdm
start_time = time.time()
NUM_EPOCHS=5

for epoch in range(NUM_EPOCHS):

    model.train()

    for batch_idx, batch in tqdm(enumerate(train_loader)):

        ### Prepare data
        input_ids = batch['input_ids'].to(DEVICE)
        attention_mask = batch['attention_mask'].to(DEVICE)
        labels = batch['labels'].to(DEVICE)

        ### Forward
        outputs = model(input_ids, attention_mask=attention_mask, labels=labels)
        loss, logits = outputs['loss'], outputs['logits']

        ### Backward
        optim.zero_grad()
        loss.backward()
        optim.step()

        ### Logging
        if not batch_idx % 250:
            print (f'Epoch: {epoch+1:04d}/{NUM_EPOCHS:04d} | '
                   f'Batch {batch_idx:04d}/{len(train_loader):04d} | '
                   f'Loss: {loss:.4f}')

    model.eval()

    with torch.set_grad_enabled(False):
        print(f'Training accuracy: '
              f'{compute_accuracy(model, train_loader,DEVICE):.2f}%'

    print(f'Time elapsed: {(time.time() - start_time)/60:.2f} min')

print(f'Total Training Time: {(time.time() - start_time)/60:.2f} min')
print(f'Test accuracy: {compute_accuracy(model, test_loader,DEVICE):.2f}%')
```

Figure 3.5: Training loop for fine-tuning process.

Experiment B: Parallel Approach

In this experiment, we attempt to evaluate the performance of the existing "fact-or-opinion-xlmr-el" model which is a multi-lingual XLM-R model fine-tuned specifically for subjectivity classification task in the English language. The model was evaluated on the testing set of our augmented data set translated from Arabic text into English text.
First, we translate the Arabic texts in our augmented data test set into English using **Googletrans**, a python library that uses Google Translate API [4]. This step allows us to bridge the language gap and make use of the fine-tuned "fact-or-opinion-xlmr-el" model, which operates on English text. The Table 3.3 includes some examples of the translated texts from the augmented dataset, providing a glimpse into the

[4] https://pypi.org/project/googletrans/

content that the "fact-or-opinion-xlmr-el" model was evaluated on during the subjectivity classification task.

Table 3.3: Examples of Arabic texts translated with GoogleTrans.

Arabic text	Translated to English	Dataset
أشارك الليلة في برنامج على قناة الناس الثامنة والنصف مساءً (عصام سلطان)	I am participating tonight in the program on Al-Nas Channel, 8:30 pm (Issam Sultan)	ASTD
فعلا نصائح مميزة ومفيدة جدا احسن حاجة في الأحلام أنها بتكتب عن تجربة	Really special and very useful tips. The best thing about dreams is that they write about experience	LABR
"فندق فاشل". انا حجزت ووصلت فالموعد ولم اجد غرف. ما يصلح فاشل النظام	Fail hotel. I booked and arrived on time but did not find rooms. What fixes the failure of the system	HARD
أعلنت شركة جوتن، إحدى أبرز الشركات العالمية في مجال إنتاج وتوريد الدهان والطلاء وبودرة الطلاء، عن إطلاقها النسخة العربية من موقعها الإلكتروني	Jotun, one of the most prominent international companies in the field of producing and supplying paints, coatings, and powder coatings, announced the launch of the Arabic version of its website	SANAD

Next, we predict the translated English text's subjectivity using the fine-tuned "fact-or-opinion-xlmr-el" model. By doing so, we aim to analyze the impact of the translation process on the quality of the predictions and understand how effectively the model could generalize from the translated Arabic text. The parallel approach process can be expressed in the diagram found in Figure 3.6.

The evaluation of the "fact-or-opinion-xlmr-el" model on the translated test set yielded notable results in terms of F1 score (86%), and other performance metrics are detailed in Table 3.7.

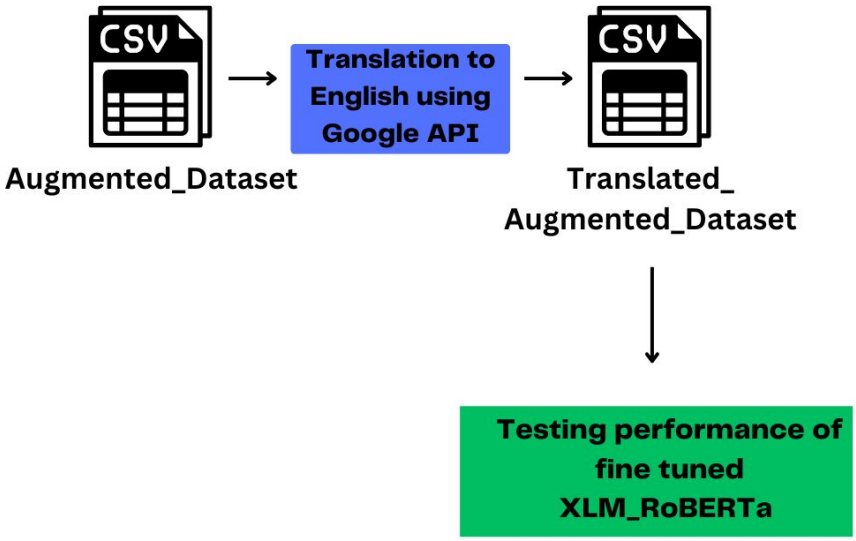

Figure 3.6: Parallel approach process.

3.6 Results and Discussion

In this part, we give a detailed presentation of the experimental results obtained in this study, revealing the findings and discussing the performances of different models. The experiments aimed to build a tool for subjectivity classification on Arabic text using different approaches and datasets. The results and discussions presented in this section shed light on the effectiveness of the fine-tuned XLM-R and AraBERT models and the potential of data augmentation for improving their performances on the task of subjectivity classification for Arabic text.

Table 3.4: Performances of AraSubjXLM-R_1 on the over-sampled ASTD.

Technique	Accuracy	Precision	Recall	F1
AraSubjXLM-R_1	82	78	89	83

The initial fine-tuning process on the over-sampled ASTD dataset (AraSubjXLM-R_1) yielded encouraging results, with the XLM-R model achieving an accuracy of 82% with the precision, recall, and F1 score mentioned in Table 3.4.

These outcomes signified the model's capacity to discern subjective and objective content. However, motivated by the pursuit of even higher performance, we recognized the potential for further improvement through data augmentation.

Table 3.5: Performances of AraSubjXLM-R_2

Technique	Accuracy	Precision	Recall	F1	Dataset
AraSubjXLM-R_2	77	77	77	77	ASTD
	99	100	99	100	LABR+HARD
	100	100	100	100	SANAD
	96	96	96	96	Augmented Data

Table 3.5 shows the results obtained of XLM-R fine tuned on the augmented data set (AraSubjXLM-R_2), this experiment demonstrates the effectiveness of the data augmentation technique in improving the performance of the model for subjectivity classification on Arabic text.

AraSubjXLM-R_2 achieved accuracy and F1-score of 96 %. These results indicate a considerable improvement in the model's performance compared to AraSubjXLM-R_1.

Furthermore, to gain deeper insights into this performance on the augmented dataset, we evaluated the performance on individual parts of testing sets that represent 20% of each dataset. The results indicate that AraSubjXLM-R_2 achieved high accuracy and F1-scores on all portions of the training set, ranging from 77% for the ASTD, 100% for SANAD, and finally 99% for both reviews (LABR and HARD).

Table 3.6: Performances of AraSubjBERT

Technique	Accuracy	Precision	Recall	F1	Dataset
AraSubjBERT	82	77	93	84	ASTD
	100	100	100	100	LABR+HARD
	100	100	100	100	SANAD
	97	95	99	97	Augmented Data

We also evaluated the performance of the Arabert model fine-tuned on the augmented data set (AraSubjBERT). The results are shown in Table 3.6, where high accuracy and F1-scores are achieved by the model (97 %) on the test set.

Similar to the evaluation of the AraSubjXLM-R_2 model, we further evaluated the performance of the AraSubjBERT model on each part of the testing set. The results indicated that the model achieved high accuracy and F1 scores on all parts of the testing set, ranging from 82% for the ASTD, 100% for SANAD, and 100% for the reviews (LABR and HARD).

Moreover, we compared the performance of the AraSubjBERT model with the AraSubjXLM-R_2 model. The results showed that AraSubjBERT outperformed AraSubjXLM-R_2 model in terms of accuracy and F1-score on the test set. However, the difference in performance was not significant, and both models achieved high accuracy and F1-scores on the test set.

The results obtained from the parallel approach experiment (Table 3.7) showcase the effectiveness of the fine-tuned XLM-R: "fact-or-opinion-xlmr-el" model in subjectivity classification task on English text that was translated from Arabic text.

Table 3.7: Performances of fact-or-opinion-xlmr-el.

Technique	Accuracy	Precision	Recall	F1	Dataset
fact-or-opinion-xlmr-el	55	59	36	44	ASTD
	90	100	90	95	LABR+HARD
	97	100	97	98	SANAD
	87	92	81	86	Augmented Data

The model achieved an accuracy of 87% on the translated test set, which is quite a promising result for the task of Arabic subjectivity classification.

Additionally, we tested the model on each portion of the testing set as we did earlier. The evaluation results indicate varying levels of accuracy, precision, recall, and F1-scores on the different data sets.

On the ASTD data set, the model achieved an accuracy of 55%, precision of 59%, recall of 36%, and F1-score of 44%. This suggests that the model had difficulty classifying this dataset's subjective and objective text. Whereas the model achieved an accuracy of 97% for SANAD and 90% for the reviews datasets (LABR and HARD).

However, it is important to note that the translation process may have somewhat impacted the quality of the predictions. This could be due to the nuances and cultural references unique to the Arabic language, which may have been lost in translation. Additionally, the quality of the translation itself could have affected the model's performance.

Another factor that can impact the performance is the existence of dialect Arabic in the text which can pose a challenge because dialect Arabic can have significant differences in grammar, vocabulary, and spelling compared to standard Arabic. Therefore, the presence of dialect Arabic in the text can affect the model's ability to accurately classify the text's subjectivity.

3.7 Conclusion

This chapter aimed to satisfy the purpose of our study, which is developing an Arabic tool for subjectivity classification by conducting various experiments on different models, including fine-tuning and evaluating the pre-trained AraBERT and XLM-R models on an augmented dataset. Additionally, we assessed the effectiveness of translating Arabic text into English and using an existing fine-tuned XLM-R model for subjectivity classification in English. Overall, the findings of this chapter suggest that subjectivity classification in Arabic text can be effectively performed using fine-tuning pre-trained language models.

Chapter 4

Corpus Construction for Arabic QA Subjectivity

4.1 Introduction

This chapter presents the detailed methodology and experiments conducted to build an Arabic question and answer (QA) corpus for subjectivity analysis. The next sections outline the process through which we obtain and analyze question-and-answer (QA) data from QA platforms. Our objective is to gather a substantial corpus of QA data and classify it into subjective and objective categories using the models mentioned in Section 3.5. By following a comprehensive approach, involving data scraping, cleaning, and classification, we aim to gain insights into the subjective and objective nature of the questions and answers.

4.2 Analysis Process

We present a comprehensive methodology for building a corpus of Arabic QA data specifically designed for subjectivity analysis. The methodology involves a multi-step process, combining Web scraping for data collection, data cleaning, and annotation process.

Web Scraping: is a technique used to automatically extract data from websites. It involves writing scripts or programs that simulate human browsing behavior to navigate Web pages, locate relevant information, and extract it for further analysis (Khder, 2021).

Figure 4.1 shows a simplified overview of how Web scraping functions and its potential applications.

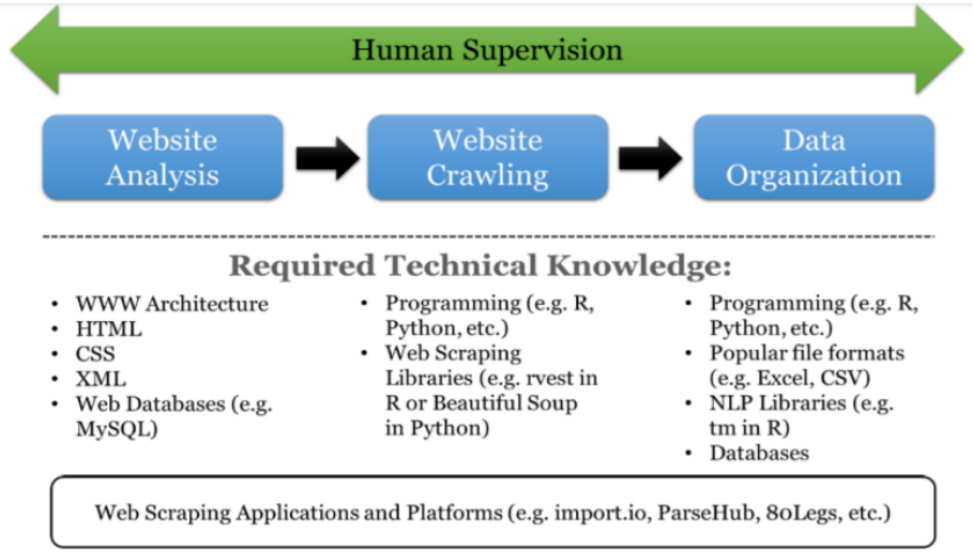

Figure 4.1: Overview of Web Scraping (Krotov & Tennyson, 2018).

In our study, we utilize Web scraping to collect a large-scale Arabic QA data set. The first step in corpus construction involved identifying suitable websites that host Arabic question-and-answer content. We focused on diverse sources of QA platforms, as they often contain rich user-generated content with subjective and objective information. Target domains are selected based on their relevance to the subjectivity aspect under investigation.

To automate the data collection process, we develop a custom Web scraping script. This script utilizes frameworks such as Beautiful Soup or Selenium to navigate through the target websites, locate QA content, and extract relevant metadata such as the question, answer, and domain. Through this automated process, we are able to gather a large quantity of Arabic QA pairs efficiently.

To ensure the quality and consistency of the scraped QA data, we perform a series of data-cleaning steps. Once the data cleaning process is completed, The clean QA data is annotated using our two fine-tune models developed earlier for the subjectivity classification task, which involves feeding the QA data into the models to analyze the text and predict the corresponding subjectivity label. This annotation process leverages the capabilities of fine-tuning models to assign subjectivity labels to each question and answer in the data set.

4.3 Development Environment

To perform our task of building the Arabic QA corpus we conduct the experiments on Google Colab described in section 3.4. Additionally, we utilize other Python Libraries for Web scrapping such as Selenium and Beautiful Soup.

Selenium: Selenium is a powerful Web browser automation tool that acts as a wrapper, allowing users to interact with Web browsers like Google Chrome, Firefox, or Internet Explorer. By writing scripts in programming languages like Python, Java, C#, or JavaScript, to control the Web browser and simulate user

interactions. This includes actions like clicking buttons, filling out forms, navigating through pages, and extracting data (Zhao, 2017).

Beautiful Soup: Beautiful Soup is a Python package designed for extracting information from HTML and XML documents through Web scraping. It offers user-friendly functions in Python, that make it easy to navigate, search, and modify the parse tree. It provides a toolkit for breaking down an HTML file and extracting the desired information using either the *lxml* or *html5lib* libraries. One of the notable features of Beautiful Soup is its ability to automatically detect the encoding of Parsing being processed and convert it to a format that can be read by the client. This eliminates the need for manual encoding detection and conversion (Zhao, 2017).

4.4 Corpus Construction Process

4.4.1 Data Collection

Our goal is to collect a QA dataset from diverse domains in the Arabic language by scraping two popular websites, *Quora* and *Hsoub*. We have chosen these platforms for their rich content of question-and-answer pairs, which offer a valuable resource.

The Arabic QA Websites

Hsoub[1]**:** Hsoub is an Arabic online platform that facilitates knowledge sharing, discussions, and information exchange including questions and answers. With a user-friendly interface and active community participation, Hsoub has become a valuable resource for Arabic-speaking users seeking information on various topics.

The grounds for selecting Hsoub are as follows:

1. Arabic Language Focus: Hsoub is specifically designed for Arabic-speaking users, making it an ideal platform for collecting Arabic QA data. The content available on Hsoub is predominantly in Arabic, ensuring that the collected data aligns with the language requirements of the corpus development.

2. Diverse Topics: Hsoub covers a wide range of topics, encompassing technology, science, arts, culture, and more. This diversity allows for the collection of QA data from various domains, ensuring the corpus represents a comprehensive range of subject matters.

3. Engaged Community: Hsoub boasts an active and engaged user community that actively participates in discussions and knowledge sharing. This engagement facilitates the availability of quality QA content, as users often provide detailed and informative responses to questions.

4. User-Generated Content: Hsoub's content is largely generated by its users, making it a valuable source for authentic user perspectives and experiences. User-generated content can provide valuable insights into subjective aspects, opin-

[1] `https://io.hsoub.com/`

ions, and diverse viewpoints, enhancing the subjectivity analysis of the QA data.

5. Accessibility and Availability: Hsoub offers a user-friendly platform that is accessible to a wide range of Arabic-speaking individuals. The ease of access and availability of content ensures a sizable pool of data for collection, increasing the likelihood of obtaining a substantial corpus.

Figure 4.2 illustrates the interface of the Hsoub website, a platform for asking and answering questions. It showcases a clean design with a question-answer display.

Figure 4.2: Hsoub interface.

Quora[2]**:** Quora is a popular global question-and-answer platform that connects users with knowledge and expertise. It covers a wide range of topics and supports a community-driven approach to sharing information and engaging in discussions.

The reasons for choosing Quora:

1. Wide User Base: Quora has a large and diverse user base, including Arabic-speaking users. This broad user community provides an opportunity to collect Arabic QA data from a wide range of individuals, ensuring the corpus represents diverse perspectives and subject matters.

2. Content Variety: Quora covers a vast array of topics, ranging from technology and science to arts, culture, and personal experiences. This diverse range of subjects enables the collection of Arabic QA data across different domains, enriching the corpus with a wide range of topics and discussions.

3. Quality Responses: Quora's platform encourages users to provide detailed and informative answers. Many users take the time to craft well-thought-out responses, making Quora a valuable source for acquiring high-quality QA data. The platform's emphasis on informative and helpful content enhances the reliability and accuracy of the collected data.

4. Linguistic Considerations: Quora supports multiple languages, includ-

[2]https://ar.quora.com/

ing Arabic, allowing for the collection of Arabic QA data in a native language context. This linguistic aspect ensures that the collected data aligns with the language requirements of the corpus development, capturing the nuances and expressions specific to Arabic.

5. Community Validation: Quora incorporates a community-driven validation mechanism where users can upvote, downvote, and comment on answers. This validation process adds a layer of quality control, as popular and well-regarded answers receive recognition from the community. This validation mechanism contributes to the reliability and accuracy of the collected QA data.

Figure 4.3 illustrates the interface of the Quora website, a platform for asking and answering questions. It showcases a clean design with a question-answer display.

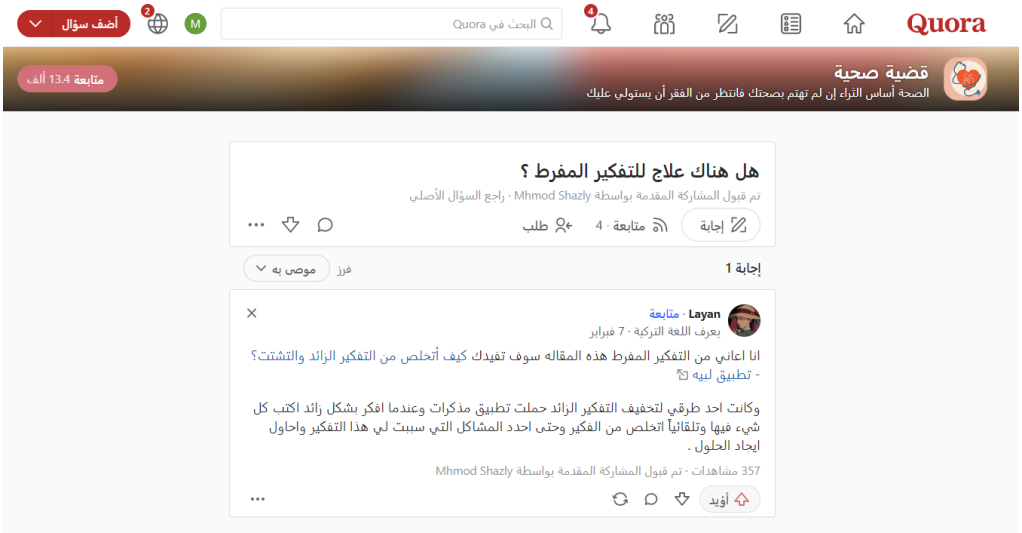

Figure 4.3: Quora interface.

The scrapping Process

We employ Web scraping techniques to navigate through the pages of the website and extract the relevant content. The process involves the following steps:

Page Navigation: We use Selenium, a powerful Web scraping tool, to navigate through Quora's and Hsoub's pages. Selenium allowed us to simulate user interactions by automatically scrolling through question lists and accessing individual question pages.

Content Extraction: Once on a question page, we use Selenium to acquire the question and its recommended answer for the Quora case. However, for Hsoub, we retrieve the question and its best response.

After the scraping process was completed, we collect a large amount of data from the chosen websites, Hsoub and Quora. From Hsoub, we scraped a total of $14,589$ question-and-answer pairs. This extensive collection provides a significant resource for our project and encompasses a various domains, including news, culture, and general knowledge. The diverse nature of the data obtained from Hsoub ensures that our dataset includes different topics and perspectives.

Additionally, our efforts on Quora yielded 3,391 question-and-answer pairs. Although the quantity may be relatively lower compared to Hsoub, Quora contributes valuable content due to its active user base and the broad array of topics discussed on the platform. The data obtained from Quora covers diverse domains and offers a rich source of information for our project.

By combining the results from both websites, we have obtained a significant total of 17,980 question-and-answer pairs in Arabic. Furthermore, it is important to note that the scraped QA data will undergo a thorough cleaning process to ensure its quality and consistency.

4.4.2 Data Cleaning

After collecting the raw QA data through Web scraping, we proceed with data cleaning steps to ensure the quality and consistency of the corpus. The cleaning process involves several key steps:

1. Duplicate Removal: We automate the elimination of duplicate question and answer pairs from the collected data. This automated process involves comparing the content of each QA pair and retaining only one instance of identical or highly similar pairs. By leveraging coding techniques, we ensure efficiency and accuracy in identifying and removing duplicate entries from the corpus.

2. Noise Filtering: We remove noisy content, in which, we apply filters to retain only the questions that contain certain keywords or question structures, such as: "اين" (where), "لماذا" (why), "ماذا"(what), and the Arabic question mark "؟", ...

By employing these data-cleaning techniques, we enhance the quality and relevance of the scraped QA data. The removal of duplicates ensures data integrity, while the application of the noise filters allows us to focus our analysis on high-quality questions. These cleaning steps pave the way for subsequent analyses, including subjectivity classification and further exploration of the subjective and objective elements within the data set.

Upon completing the cleaning process of the QA data collected from the Hsoub and Quora websites, we have successfully obtained a refined and high-quality corpus of Arabic QA pairs. From the Hsoub website, the clean dataset consists of 9,662 Arabic QA pairs, while the clean dataset from Quora comprises 2,735 QA pairs (Table 4.1).

Table 4.1: Size of datasets before and after cleaning.

Dataset	**Size before cleaning**	**Size after cleaning**
Hsoub	14,589	**9,662**
Quora	3,391	**2,735**
Total dataset	17,980	**12,397**

4.4.3 Data Annotation

After the data cleaning process, the next step in our methodology involves annotating the cleaned QA data for subjectivity using our two fine-tuned models, AraSubjXLM-R_2 and AraSubjBERT. These models have been trained on our Augmented Arabic language dataset and fine-tuned on subjectivity classification tasks, allowing them to accurately predict the subjectivity label of each question and answer in the QA dataset.

The annotation process involves feeding the clean QA data into the AraSubjXLM-R_2 and AraSubjBERT models, which analyze the text and predict the corresponding subjectivity label for each question and answer. Leveraging the knowledge and insights gained from the fine-tuning process, the models determine whether the content is subjective or objective.

The combination of two separate models also provides an additional layer of validation for the subjectivity annotations. By comparing the subjectivity predictions made by both models, in cases where there is a disagreement between the models, we eliminate that pair of Question and answer and keep only the QA pair where the models agree so we can ensure the reliability and consistency of the annotations throughout the dataset.

After applying the elimination process based on agreement between the AraSubjXLM-R_2 and AraSubjBERT models, our corpus consists of 6,293 QA pairs (4,805 from Hsoub and 1,488 from Quora), instead of the initial 9,662 pairs (Table 4.2).

Overall, by annotating the clean QA data using these two fine-tuned models, we can generate a comprehensive and accurate classification of subjectivity for each question and answer in the corpus. This annotated corpus provides a valuable resource for subsequent analyses and insights into the subjective and objective aspects of the QA content.

Table 4.2: QA dataset size before and after annotation.

Dataset	**Size before annotation**	**Size after annotation**
Hsoub	9,662	**4,805**
Quora	2,735	**1,488**
Total dataset	12,397	**6,293**

4.4.4 Data Format

The Arabic QA corpus files are structured in a widely used and easily accessible format known as CSV (Comma-Separated Values). Each file within the corpus is dedicated to a specific domain, ensuring that the data is organized and categorized according to relevant subject areas. In terms of the file structure, each CSV file is organized into rows and columns, with each row representing a distinct question-answer pair and each column representing a specific attribute as follows:

Question: The question is extracted from the scrapped Arabic websites (Hsoub or Quora) after cleaning the irrelevant text.

Answer: the corresponding answer to the question posed (Best answer in Hsoub, recommended answer in Quora).

Domain: the scrapped spaces from Quora such as:(أخبار , ثقافة عالمية , قضية صحية...) and communities from Hsoub (تسلية , العمل الحر , تطوير الواب ...).

Question Annotation 1: the Question subjectivity label predicted by AraSubjXLM-R_2.

Question Annotation 2: the Question subjectivity label predicted by AraSubjBERT.

Question Final Annotation: the agreed subjectivity label.

Answer Annotation 1: the Answer subjectivity label predicted by AraSubjXLM-R_2.

Answer Annotation 2: the Answer subjectivity label predicted by AraSubjBERT.

Answer Final Annotation: the agreed subjectivity label.

4.5 Results and Discussion

In this section, We showcase examples of the constructed corpus, providing insights into the nature of the collected QA pairs. Additionally, we discuss relevant statistics to shed light on the size, diversity, and subjectivity distribution within the corpus. Also providing a foundation for further analysis and evaluation of subjectivity in Arabic text.

Table 4.3 provides a glimpse of the diverse nature of our QA corpus and showcases a selection of question-answer pairs along with their respective subjectivity labels.

Moreover, The examples within the corpus highlight the subjectivity present in the corpus, allowing for an examination of the relationship between the subjectivity of the question and its corresponding answer. This analysis provides a deeper understanding of how subjectivity is conveyed and perceived in Arabic QA interactions. By utilizing this corpus, researchers and practitioners can enhance their NLP applications and contribute to advancing the understanding of subjectivity in textual data.

Table 4.3: Examples of the AQA subjectivity corpus.

Question	Q final annotation	Answer	A final annotation
هل نستطيع بالحفارات التي لدينا اليوم أن نبني موطن داخل الجبال الصخرية كما بنت ثمود مواطنها فيها؟	1	يمكننا ذلك لكن بالعقل بحفارات أو بدون حفارات لكن أعتقد أننا لا نستطيع أن نبني حضارة كاملة مثل حضارتهم لأن الله وصفها بوصف أنها لم يخلق مثلها فالبلاد والله أعلى وأعلم	1
هل الافضل افعل تمارين التمدد و انا صائم ام افعلها بعد الفطور؟	0	يفضل أن تقوم بتمارين التمدد بعد الفطور، حيث يحتاج جسمك إلى الطاقة والتغذية للقيام بالتمارين بشكل صحيح وآمن. كما أنه من المهم شرب الماء بشكل كاف خلال فترة الصيام وبعدها لتجنب الجفاف وضربات الشمس	0
ما هي لغة البرمجة التي يجب تعلمها للعمل في تطوير الذكاء الاصطناعي؟	0	الافضل في مجال الذكاء الصناعي python	1
ما هي فوائد أُوميجا ٣؟	0	تعزيز صحة القلب أثبتت العديد من الأبحاث أن زيادة استهلاك زيت السمك تقلل من عوامل الخطر للإصابة بأمراض الأوعية الدموية في القلب - خفض مستويات الدهون الثلاثية - تقليل خطر الولادة المبكرة	0

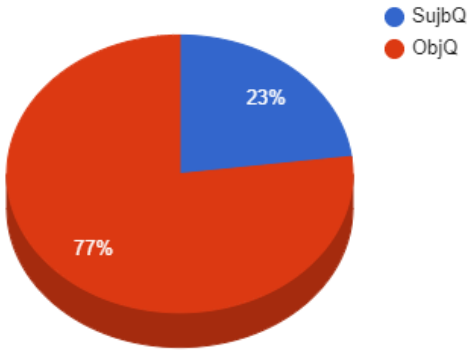

Figure 4.4: Percentage of subjectivity in questions.

Figure 4.4 shows that 77 % of the questions are classified as objective. This suggests that a significant portion of the questions seek factual or objective information rather than personal opinions or subjective perspectives. Users may be looking for concrete answers, definitions, or specific details when posing these objective questions.

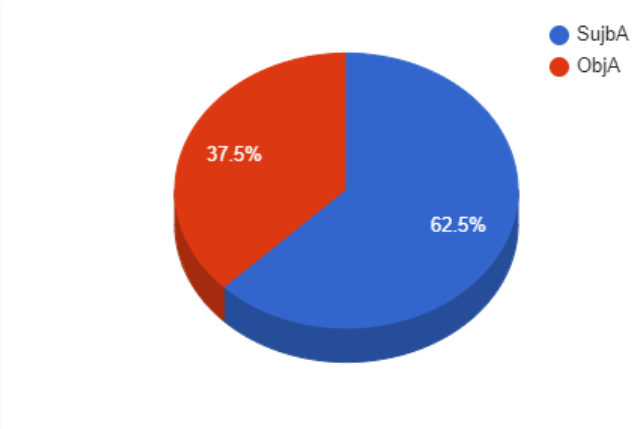

Figure 4.5: Percentage of subjectivity in answers.

The observation in Figure 4.5 reveals that 62.5 % of the answers are categorized as subjective indicates a higher prevalence of subjective responses provided by users. This suggests that the community actively engages in sharing their personal opinions, interpretations, or subjective knowledge when responding to questions on the Arabic platforms.

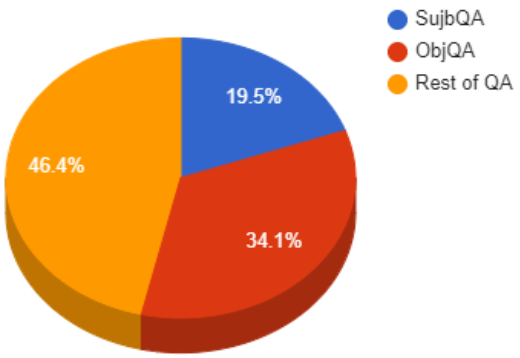

Figure 4.6: Percentage of subjectivity distribution in QA.

Furthermore, we can conclude from the Figures 4.4 and 4.6 that 85% of the cases where the question is subjective, the answer is also subjective. This highlights a strong correlation between the subjectivity of the question and the subjectivity of the corresponding answer. This indicates that when users pose subjective questions, they are more likely to receive subjective responses from the community.

While, within the objective questions, approximately 44.3% of the corresponding answers are also categorized as objective. This indicates that while the fact that objective questions receive objective answers, a significant portion of them may still get subjective responses. It suggests that even within objective inquiries, there might be room for individual interpretation, personal experiences, or varying viewpoints.

4.6 Conclusion

This chapter provides a comprehensive methodology for building an Arabic question and answer (QA) corpus for subjectivity analysis. The methodology involves Web scraping techniques to collect QA pairs from Hsoub and Quora websites. The cleaned QA pairs are then annotated using fine-tuned models, AraSubjXLM-R_2 and AraSubjBERT, to classify them based on subjectivity. Overall, the successful completion of these experiments highlights the importance of collecting and cleaning high-quality data for building effective models and conducting accurate analysis. The resulting Arabic QA corpus provides a valuable resource for studying subjective aspects of Arabic text, enabling researchers to gain insights into user opinions, sentiments, and subjective perspectives.

Conclusion and Perspectives

This study addressed the scarcity of resources in Arabic subjectivity analysis by constructing a valuable Question-Answer (QA) corpus focused on subjectivity classification. The creation of this corpus contributes significantly to the limited resources available for Arabic subjectivity analysis and sentiment classification.

The construction and annotation process of the Arabic QA corpus involved several key steps, each playing a crucial role in the process. First and foremost, we dedicated significant effort to developing and employing subjectivity classification models by fine-tuning two existing pre-trained models. These models were trained using our augmented dataset. Furthermore, we collect a diverse and representative corpus, ensuring that it encompasses a wide range of subjective expressions. Next, meticulous cleaning procedures were carried out to remove noise and irrelevant information, ensuring the integrity and quality of the corpus.

We successfully annotated the AQA corpus with subjectivity labels using our two developed models. The constructed and annotated Arabic QA corpus along with the two developed models for subjectivity analysis serve as crucial resources for researchers and practitioners in Arabic NLP. It enables training and evaluation of subjectivity analysis models, benchmarking of algorithms and techniques, and facilitates advancements in sentiment analysis, opinion mining, and social media analytics for Arabic.

While this study has made significant contributions to Arabic subjectivity analysis through the construction of a Question-Answer (QA) corpus for subjectivity, It is important to recognize its limitations. Firstly, the effectiveness of the subjectivity classification models depends on the quality of the labeled datasets used for fine-tuning. Additionally, the construction of the Arabic QA corpus relies on scraping QA pairs from platforms, which may still contain noise or irrelevant information in the collection even after the cleaning. Lastly, the automated approach used for label assignment may not capture the subjective information accurately.

In terms of perspectives, there are several avenues for further exploration and improvement in Arabic subjectivity analysis. One important aspect is expanding our augmented data set by incorporating more diverse data sources and the inclusion of dialectal Arabic. By incorporating dialectal Arabic data, researchers can gain insights into subjective language use across different dialects and enhance the models' ability to handle dialectal variations. Another perspective is the utilization of manual annotation to evaluate the performance of subjectivity classification models. While the use of automated approaches for annotation provides efficiency, manual annotation by human annotators can offer a more accurate and nuanced assessment of subjectivity.

Bibliography

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM transactions on information systems (TOIS)*, *26*(3), 1–34.

Abdul-Mageed, M., Diab, M., & Korayem, M. (2011). Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 587–591).

Abdul-Mageed, M., Diab, M., & Kübler, S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, *28*(1), 20–37.

Abdul-Mageed, M., & Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Lrec* (Vol. 515, pp. 3907–3914).

Abdul-Mageed, M., & Diab, M. T. (2014). Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *Lrec* (pp. 1162–1169).

Alduailej, A., & Alothaim, A. (2022). Araxlnet: pre-trained language model for sentiment analysis of arabic. *Journal of Big Data*, *9*(1), 1–21.

Alharbi, A., Kalkatawi, M., & Taileb, M. (2021). Arabic sentiment analysis using deep learning and ensemble methods. *Arabian Journal for Science and Engineering*, *46*, 8913–8923.

Alhumoud, S. O., & Al Wazrah, A. A. (2022). Arabic sentiment analysis using recurrent neural networks: a review. *Artificial Intelligence Review*, *55*(1), 707–748.

Allam, A. M. N., & Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, *2*(3).

Alotaibi, S. S. (2016). Sentiment analysis in arabic: An overview. *International Journal of Sciences: Basic and Applied Research*, *26*(2).

Al-Rubaiee, H., Qiu, R., & Li, D. (2016). Identifying mubasher software products through sentiment analysis of arabic tweets. In *2016 international conference on industrial informatics and computer systems (ciics)* (pp. 1–6).

Aly, M., & Atiya, A. (2013). 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)* (pp. 494–498).

Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Aouichat, A., & Guessoum, A. (2017). Building talaa-afaq, a corpus of arabic factoid question-answers for a question answering system. In *Natural language processing and information systems: 22nd international conference on applications of natural language to information systems, nldb 2017, liège, belgium, june 21-23, 2017, proceedings 22* (pp. 380–386).

Awwad, H., & Alpkocak, A. (2016). Performance comparison of different lexicons for sentiment analysis in arabic. In *2016 third european network intelligence conference (enic)* (pp. 127–133).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chen, H., Sun, M., Tu, C., Lin, Y., & Liu, Z. (2016). Neural sentiment classification with user and product attention. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1650–1659).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duwairi, R., & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for arabic text. *Journal of Information Science*, *40*(4), 501–513.

Einea, O., Elnagar, A., & Al Debsi, R. (2019). Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in brief*, *25*, 104076.

El Karfi, I., & El Fkihi, S. (2022). An ensemble of arabic transformer-based models for arabic sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *13*(8).

Elnagar, A., Khalifa, Y. S., & Einea, A. (2018). Hotel arabic-reviews dataset construction for sentiment analysis applications. *Intelligent natural language processing: Trends and applications*, 35–52.

Ghosal, D., Akhtar, M. S., Chauhan, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3454–3466).

Giannakopoulos, A., Musat, C., Hossmann, A., & Baeriswyl, M. (2017). Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets. *arXiv preprint arXiv:1709.05094*.

Google colab. (2023). Retrieved from `https://research.google.com/colaboratory/faq.html`

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-w. (2020). Realm: Retrieval-augmented language model pre. *Training*.

Ismail, W. S., & Homsi, M. N. (2018). Dawqas: A dataset for arabic why question answering system. *Procedia computer science*, *142*, 123–131.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … others (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, *13*(3).

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the international aaai conference on web and social media* (Vol. 5, pp. 538–541).

Krotov, V., & Tennyson, M. (2018). Research note: Scraping financial data from the web using the r language. *Journal of Emerging Technologies in Accounting*, *15*(1), 169–181.

Liu, B., et al. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*(2010), 627–666.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093–1113.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (icics)* (pp. 243–248).

Mountassir, A., Benbrahim, H., & Berrada, I. (2012). A cross-study of sentiment classification on arabic corpora. In *Research and development in intelligent systems xxix: Incorporating applications and innovations in intelligent systems xx proceedings of ai-2012, the thirty-second sgai international conference on innovative techniques and applications of artificial intelligence* (pp. 259–272).

Nabil, M., Aly, M., & Atiya, A. (2015). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2515–2519).

Numpy documentation. (2023). Retrieved from `https://numpy.org/doc/stable/#`

pandas. (2023). Retrieved from `https:https://pypi.org/project/pandas/`

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825-2830. Retrieved from `http://jmlr.org/papers/v12/pedregosa11a.html`

Python official website documentation. what is python? executive summary. (2023). Retrieved from `https://www.python.org/doc/essays/blurb/` (Accessed on: 13/05/2023)

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ranasinghe, T., & Zampieri, M. (2020). Multilingual offensive language identification with cross-lingual embeddings. *arXiv preprint arXiv:2010.05324*.

Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1650–1659).

Rushdi-Saleh, M., Martín-Valdivia, M. T., Lopez, L. A. U., & Perea-Ortega, J. M. (2011). Bilingual experiments with an arabic-english corpus for opinion mining. In *Proceedings of the international conference recent advances in natural language processing 2011* (pp. 740–745).

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., … others (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shaheen, M., & Ezzeldin, A. M. (2014). Arabic question answering: systems, resources, tools, and future trends. *Arabian Journal for Science and Engineering*, *39*, 4541–4564.

Tomás, D., Vicedo, J. L., Bisbal, E., & Moreno, L. (2009). Trainqa: a training corpus for corpus-based question answering systems. *Polibits*(40), 5–11.

Torch. (2023). Retrieved from `https://pytorch.org/docs/stable/torch.html`

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, B., Spencer, B., Ling, C. X., & Zhang, H. (2008). Semi-supervised self-training for sentence subjectivity classification. In *Advances in artificial intelligence: 21st conference of the canadian society for computational studies of intelligence, canadian ai 2008 windsor, canada, may 28-30, 2008 proceedings 21* (pp. 344–355).

Wang, W., Pan, S. J., Dahlmeier, D., & Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Wettig, A., Gao, T., Zhong, Z., & Chen, D. (2022). Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … others (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wu, Y., Zhang, Q., Huang, X.-J., & Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 1533–1541).

Zhao, B. (2017). Web scraping. *Encyclopedia of big data*, 1–3.