الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
جامعة عمّار ثليجي بالأغواط
AMAR TELIDJI LAGHOUAT UNIVERSITY

كلية العلوم
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS AND INFORMATICS

Master

Domaine : Mathematics and Computer science
Filiere : **Computer science**
Option : System Informatics and Decision

by :
Rezig Afifa Aida
Bekhelifa Ikram

Theme

Towards a Dataset for Arabic Multimodal Sentiment Analysis

Supervised by: Hadda Cherroun

Ouinten Youcef	Professor	President
Benameur Ziani	MCA	Examinator
Chellama Laradj	MAA	Examinator
Cherroun Hadda	Professor	Encadreur

Academic Year 2022/2023

Acknowledgements

We are with great pleasure that we reserve these few lines as a sign of gratitude to all those who have contributed directly or indirectly to the development of this work.

First of all, we would like to address our most sincere thanks to our supervisor Mme. Cherroun Hadda for her availability, patience, and precious follow-up throughout the realization of this work.

We would also like to thank the members of the jury for having devoted part of their time to reading this thesis and for their interest in this work.

A special thank you to Mr. Haouhat for your exceptional guidance and support. Your wealth of experience has been instrumental in our accomplishments.

Our thanks extend to all our teachers in the Computer Science department of the Amar Telidji The University of Laghouat.

Finally, we would like to thank all the people who contributed directly or indirectly to the accomplishment of this work.

Rezig & Bekhlifa.

شُـــكْـــر وَ إِهـــدَاء

الحَمدُ لِلَه الذي يسر البدايات وأكمل النهايات وبلغنا الغايات

الحمد لله الذي بنعمته تتم الصالحات

الحمد لله الذي ما تم جهد إلا بعونه وما ختم سعي إلا بفضله

الحمد لله على البلوغ ثم الحمد لله على التمام

الحمد لله الذي ماتيقنت به خيراً وأملاً إلا وأغرقني سروراً.

إلى أسمى آيات العطاء البشريّ، أمي وأبي الغاليين،

أهدي ثمرة جهدي المتمثلة في هذا البحث المتواضع،

عسى أن أكون مصدر فخر لكما.

إلى من مَدّتن بأياديهن في أوقات الضعف، غير راضيات باستكانتي

إليكن يا أخواتي الدّاعمات لي في

أحلك الظروف أهدي اليكن لحظة فرحي.

عَـفِـيـفَـة عـايْـدَة رزِيـقْ

Dedication

In the Name of Allah the Most Merciful, the Most Compassionate

First and foremost, I must acknowledge my limitless thanks to Allah, for His help and bless by giving me the opportunity, courage and enough energy to carry out and complete the entire thesis work.

To the first two heroes in my life, to those who sacrificed their lives for mine, to those who paved the way for me at the expense of their happiness and comfort, to those whom I have not done enough for, to those who inspired me with their strength and patience, to the two individuals i am most proud of, to the two people who are the reason i have reached this point — My beloved parents. A simple "thank you" is far less than what you deserve, but i will express my gratitude nonetheless. Thank you for being the shelter and home i could always rely on. Thank you for silently bearing so much. Thank you for supporting me even when i didn't deserve it. Simply put, thank you for being my parents.

To my first team in life, to those who have been my crutch during my moments of brokenness, to my half, my siblings, Fatima Zohra, Amina, Ayat. Thank you for being there for me during my times of weakness, and thank you for continuing to stand by me in my moments of strength now. To the one for whom I have exerted all my efforts, so that she will be proud of me in the future, my second daughter, Aicha. Thank you for being a part of my life.

To the person who unwaveringly believed in me when no one else did, to the one who saw my strength before i could see it, to the first person who supported me on this journey, Aouissi Amine. Your support and faith in me have been the driving force behind my accomplishments. Thank you. I am forever grateful for your presence in my life.

To those who shared the journey with me through its ups and downs, to my best friends and sisters, Zineb, Khadra, Doha, Amina Djoudi. Thank you for your support. Your honesty and genuine support have been invaluable to me.

And finally, to the one who deserves the most thanks and love, to myself. Thank you for believing in me. Thank you for overcoming everything that has come my way. I am extremely proud of you, you did great! Yet, this is only the beginning.

Graduation day has arrived ...

Finally i can confidently proclaim ...

I did it .

Bekhelifa Ikram.

Abstract

All around the world, Individuals are consistently imparting their insights, stories, and audits through different social media platforms. Concentrating on feeling and subjectivity in these assessment recordings is encountering a developing consideration from the scholarly community and industry.

While sentiment analysis has been successful in different languages, it is now an understudied research question for videos and multimedia content in Arabic. The greatest mishaps for concentrating on this path are the absence of a legitimate dataset, technique, baselines, and factual investigation of how data from various methodology sources connect with one another.

In summary, this thesis introduces the AMMD dataset, a valuable resource for sentiment analysis in Arabic. It also, encapsulates the essence of the research, highlighting the importance of a multimodal approach and the comprehensive methodology employed for data collection and alignment. The results presented in this thesis offer insights into sentiment analysis in Arabic and pave the way for further advancements in this field. The AMMD dataset is expected to facilitate research and development of sentiment analysis applications for the Arabic language.

Keywords: Arabic sentiment analysis, Multimodal dataset, Arabic dialects.

Résumé

Partout dans le monde, les individus communiquent constamment leurs opinions, histoires et critiques via différentes plateformes de médias sociaux. La focalisation sur les émotions et la subjectivité dans ces vidéos d'évaluation suscite un intérêt croissant de la part de la communauté universitaire et de l'industrie.

Bien que l'analyse des sentiments ait connu un succès dans différentes langues, elle est actuellement une question de recherche peu étudiée pour les vidéos et le contenu multimédia en arabe. Les plus grandes lacunes pour se concentrer sur cette voie sont l'absence d'un ensemble de données légitime, de techniques, de références et d'une enquête factuelle sur la manière dont les données provenant de différentes sources méthodologiques sont liées les unes aux autres.

En résumé, cette memoire présente l'ensemble de données AMMD, une ressource pour l'analyse des sentiments en arabe. Elle encapsule également l'essence de la recherche, mettant en évidence l'importance d'une approche multimodale et de la méthodologie complète utilisée pour la collecte et l'alignement des données. Les résultats présentés dans cette memoire offrent un aperçu perspicace de l'analyse des sentiments en langue arabe et ouvrent la voie à de nouvelles avancées dans ce domaine.

Mots-clés: Analyse des sentiments arabes, Jeu de données multimodal, Dialectes arabes.

ملخص

في جميع أنحاء العالم ، ينقل الأفراد باستمرار رؤاهم وقصصهم ومراجعاتهم من خلال منصات وسائط اجتماعية مختلفة. التركيز على الشعور والذاتية في تسجيلات التقييم هذه يواجه اهتمامًا متزايدًا من المجتمع العلمي والصناعة.

بينما نجح تحليل المشاعر في لغات مختلفة، إلا أنه الآن يُعَدُّ سؤال بحثيًا غير مدروس بشكل كافٍ بالنسبة لمقاطع الفيديو والمحتوى المتعدد الوسائط باللغة العربية. أكبر الأخطاء في التركيز على هذا المسار هي غياب مجموعة بيانات مشروعة، وتقنية، وقواعد مرجعية، وتحليل حقيقي لكيفية ترابط بيانات من مصادر منهجية مختلفة.

باختصار ، تقدم هذه الأطروحة مجموعة بيانات AMMD ، وهي مورد لتحليل المشاعر باللغة العربية. كما أنه يلخص جوهر البحث ، ويسلط الضوء على أهمية النهج متعدد الوسائط والمنهجية الشاملة المستخدمة لجمع البيانات والمواءمة. النتائج المعروضة في هذه الأطروحة تقدم نظرة ثاقبة لتحليل المشاعر باللغة العربية وتمهد الطريق لمزيد من التقدم في هذا المجال.

الكلمات المفتاحية: تحليل المشاعر العربية، مجموعة بيانات متعددة الوسائط، اللغة العربية

Contents

List of Figures

List of Tables

Introduction

Sentiment Analysis (SA) is a Natural Language Processing (NLP) technique that involves the automated extraction of the sentiments, emotions, and attitudes expressed in a piece of content. This content can be a simple text, an audio, or more complex a video. The use of SA is becoming increasingly popular in fields such as marketing, customer service, and social media analysis. With this approach it is possible to gain a deeper understanding of customer sentiment and feedback, which can help businesses and organizations improve their products, services, and overall customer experience [2, 18].

Sentiments and emotions can be extracted from various modalities, such as text, audio, and visual [6]. Research has shown that a more accurate understanding of the sentiment can be obtained by considering all these factors together which improves the accuracy and completeness of SA. This process is known as Multimodal Sentiment Analysis (MSA) [2, 27, 17].

The field of Arabic Multimodal Sentiment Analysis (AMSA) has not been widely explored thus far compared to western languages. The majority of research on Arabic multimodal sentiment analysis has focused on only one or two modalities (e.g., text and video) at a time [2, 7, 15, 14]. However, no prior work has dealt with the extraction of features and integration of information extracted from all three modalities together in Arabic (i.e., text, audio, and video). Arabic is a complex language with unique cultural and linguistic characteristics. Arabic is spoken across a wide geographic region and can vary significantly across different dialects. The meanings of Arabic words can be ambiguous, and the context in which they are used can greatly affect the sentiment expressed. Combining data from various modalities requires specialized techniques for integration, and Arabic language data have different formats and structures that make integration particularly challenging. Compared to other languages, its data is relatively scarce. In addition, much of the available Arabic language data are not labeled for Multimodal sentiment analysis, which makes it difficult to train machine learning models for AMSA.

In this thesis, we design a pipe to collect and process a Multimodal dataset for Arabic sentiment analysis purpose. That integrates three modalities at a time. This dataset has been built by collecting data and then preparing it to extract relevant features in each modality, and aligning it to ensure that the data extracted are in a consistent and compatible format to get annotated. The main objective of the investigation is to build a dataset that leverages the detection of semantic information inherent in the gathered modalities. Indeed, we investigate the usage of transformers and advanced audio-visual feature extraction tools and techniques.

Besides a general introduction and conclusion, our master thesis is organized into three chapters:

- Chapter one: Generalities and Related Work. It contains a comprehensive overview of the existing research and studies related to our topic.

- Chapter two: Dataset construction. It shows our methodology for building a multimodal dataset in Arabic.
- Chapter three: Dataset Assessment. In this chapter, we will provide a detailed description of the dataset we have created. We will introduce the selected model capable of handling multimodal modalities. We will interpret the training results obtained using the model and evaluate its performance using various metrics.

Chapter 1

Generalities and Related Work

In this chapter, we provide a comprehensive overview of what is related to our topic with a review of the existing research on it, summarize key findings.

1.1 Sentiment Analysis and Variants

> *Sentiment Analysis (SA), also known as opinion mining, is a Natural Language Processing (NLP) technique that involves the automated extraction of subjective information and identifying positive, negative, or neutral sentiment, emotions, and attitudes expressed in a piece of content[23].*

Sentiment can be extracted from different types of modalities of input, including text, audio, and visual data [6]. In addition, it can be extracted from two or more modalities. For that, according to the basic modalities taken, Sentiment Analysis has two main variants: Unimodal and Multimodal. Figure1.1 illustrates those variants.

1.1.1 Unimodal

Unimodal SA focuses on extracting sentiment from a single specific modality. Each modality will be explained in detail in what follows.

Text-based Sentiment Analysis

Text sentiment analysis involves analyzing sentiment in written text, it can be categorized into three levels of detail, including :

- Document-Level: Document-level sentiment classification considers the entire document as the fundamental unit of information, with an emphasis on a particular subject matter or object. It involves analyzing the sentiment or emotion tokens expressed throughout the document as a whole, aiming to categorize it as positive, negative, or neutral based on the overall sentiment conveyed, providing insights into the sentiment expressed in the document [6].
- Sentence-level: At the sentence level, sentiment analysis performs a similar task as document-level analysis, but with a key distinction. Instead of considering the entire document as a unit, sentence-level analysis focuses on individual sentences. It allows for a more nuanced interpretation of sentiment variations throughout the text, offering a more in-depth understanding of sentiment and a finer level of detail in sentiment analysis [6].

- Aspect-level: Aspect-level sentiment analysis, alternatively referred to as feature-based or entity-based sentiment analysis. It is a more specialized level of sentiment analysis that focuses on identifying aspects or entities within a sentence and classifying them as positive or negative. Unlike document-level or sentence-level sentiment analysis, which looks at the overall sentiment, aspect-level analysis delves into the sentiment expressed towards specific aspects or features mentioned in the text. These aspects can be either explicitly stated within the sentence or implicitly inferred from the expressions of sentiment [6].

Audio-based Sentiment Analysis

Audio sentiment analysis focuses on understanding the sentiment or emotions expressed in different types of audio data, such as speech recordings, audio clips, or audiovisual content. It involves processing and analyzing various acoustic features of the audio, including tone, pitch, intensity, voice quality, and speech patterns [12]. In Speech-based Sentiment Analysis, the focus is on understanding the emotional content conveyed through the spoken words, tone of voice, intonation, and other acoustic cues where the analysis aims to identify and classify the sentiment.
Automatic Speech Recognition (ASR) technologies are widely utilized in speech-based sentiment analysis to convert spoken language into a textual form, which in turn allowed for further analysis and application of sentiment analysis methods [20, 6]. By leveraging ASR, researchers were able to effectively process and interpret spoken language, opening up opportunities for sentiment analysis on audio data. The accurate conversion of speech to text through ASR is a fundamental step that enabled subsequent techniques, such as Keyword Spotting (KWS) or feature extraction using the Conventional Frequency Cepstral Coefficients (MFCC), to be applied for sentiment analysis purposes. In essence, ASR served as the foundation that enabled these approaches to effectively analyze sentiment in speech-based data [6].

Visual-based Sentiment Analysis

Visual sentiment analysis is an emerging field of research that aims to extract sentiment-related information from visual content, as opposed to relying solely on textual data like traditional sentiment analysis. It encompasses non-verbal communication, utilizing physical actions and expressions to convey information. This form of communication includes various elements such as facial expressions, body language, gestures, eye movements, touch, and spatial usage, all of which provide insights into emotional states and sentiment orientations of visual content [12].
Researchers employ advanced techniques like Convolutional Neural Networks (CNNs) to extract meaningful visual features for analyzing visual sentiments. Additionally, Visual Sentiment Ontology (VSO) serves as a valuable resource, enabling a more nuanced analysis of visual sentiment by considering the intricate relationships between different visual cues and sentiment orientations. Visual sentiment analysis focuses on understanding and interpreting emotions, attitudes, and opinions conveyed through visual elements such as images, videos, and visual gestures, aiming to bridge the gap between verbal and non-verbal communication modalities [20, 6].

1.1.2 Multimodal

As technology continues to progress, people are increasingly using audio and visual modalities to express their ideas and emotions.

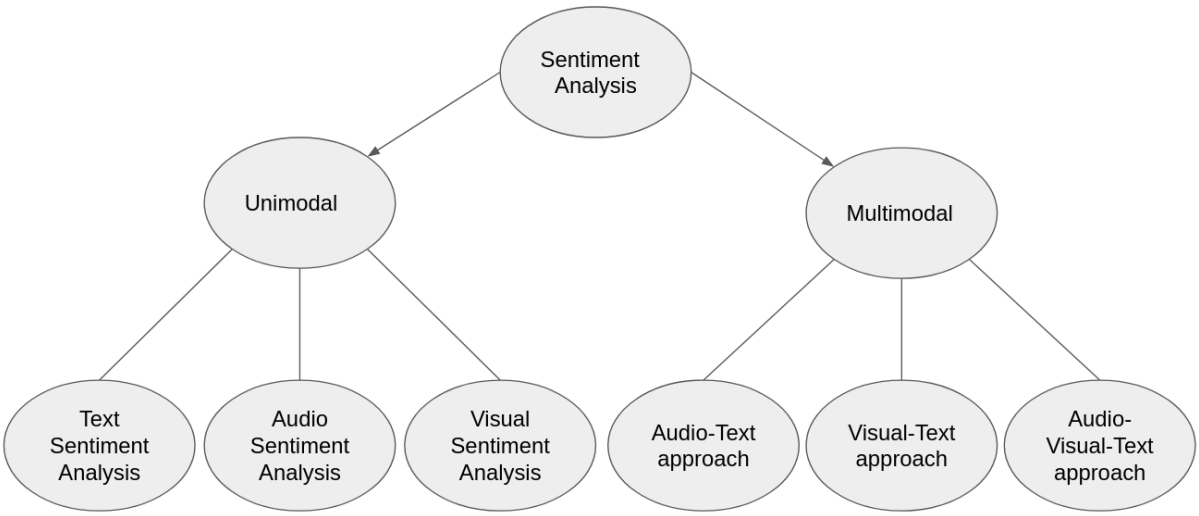


Figure 1.1: Variants of sentiment analysis.

In Multimodal sentiment analysis enhances the traditional Unimodal approach by combining and integrating data from different modalities, including text, audio, and visual content, thereby adding an extra dimension to the analysis [6].

Many issues are considered when dealing with MSA :

1. How to fuse features from different modalities.
2. How to align features extracted from different modalities.
3. How to uniformly annotate various modalities.

1.1.3 Applications

Sentiment Analysis with its variants has diverse applications across various domains, such as:

- Marketing Strategies: it plays a crucial role in understanding and analyzing customer demands and summarizing reviews and customer opinions, which in turn helps organizations enhance innovation, customer retention, and operational efficiency. This allows for the aggregation of opinions on a large scale and provides immediate feedback at a low cost. Prior to the advent of sentiment analysis, companies had to rely on surveys or focus groups, which were significantly slower and more expensive. With the rise of multimedia content on social media platforms, such as spoken reviews on YouTube, sentiment analysis has the potential to become a crowdsourced and cost-effective endeavor [12].
- Government and Politic: governments and politicians also utilize sentiment analysis to understand public sentiment toward themselves and their policies. It helps contextualize user likes and dislikes, allowing for a better understanding of people's opinions [12].
- Broadcast video news: sentiment analysis can be applied in the domain of broadcast video news, enabling automatic analysis of sentiment within news content. It can even assist in

identifying politically persuasive content , which refers to content that aims to influence or sway public opinion in favor of a particular political stance, party, or ideology. By analyzing sentiments expressed in news broadcasts across different channels and online platforms, it becomes easier to detect patterns, trends, and the effectiveness of political messaging [20, 12].

- The e-learning and e-education: Sentiment Analysis can improve the domains of e-learning and e-education by yielding valuable insights on learners' performance and providing a deeper understanding of their sentiments, opinions, and satisfaction levels [16].

1.2 Sentiment Analysis Datasets and corpus

> *A dataset is a collection of data that is treated as a single unit by a computer. This means that a dataset contains a lot of separate pieces of data and can be used to train a machine learning algorithm with the goal of finding predictable patterns inside the whole dataset [21].*

The performance of Machine learning models heavily depends on the quality of the training dataset [8]. Certainly, there are a lot of criteria to define a good data set and ensure its quality and usefulness, such as follows:

- Relevance: The data set has to capture the characteristics of the investigated property [19].
- Non-redundancy: means excluding overlapping cases within each data set [19].
- Consistency: which indicates ensuring uniform values and formats across the data set.
- Accuracy: which ensures that the data is free from errors or inconsistencies
- Scalability: it should be possible to test systems of different sizes [19].

Data set annotation is also considered as an important process in machine learning and data analysis that involves adding labels or annotations to data points in a dataset.
There are various methods employed for dataset annotation, depending on the nature of the data and the specific requirements of the task. Such as

- Manual annotation: which is the traditional approach, refers to the process of labeling or tagging data by human annotators.
- Crowd-sourcing: involves distributing the annotation task to a large group of individuals: often through online platforms.
- Automatic annotation: relies on algorithms or tools to automatically assign annotations and etc.

For the purpose of multimodal sentiment analysis, numerous datasets are available to use. According to our work focus, we can review them considering the language they support. Table 1.1 summarizes some details on the reviewed datasets.

Non-Arabic Standard Datasets

- CMU-MOSI is a collection of 2199 opinion video clips each annotated with sentiment in the range [-3,3] [26]. In fact, more fine sentiment analysis is considered.
- CMU-MOSEI (Multimodal Sentiment-Emotion Intensity) data set is the new generation of the CMU-MOSI data set. It is a large-scale multimodal data set developed by Carnegie Mellon University for sentiment analysis and emotion recognition tasks [27].
- MOUD (Multimodal OpinionUtterances Dataset), a collection of 412 utterances, 182 of which are labeled as positive, and 231 are labeled as negative. It is made of product review videos in Spanish. Each video consists of multiple segments labeled to display positive, negative, or neutral sentiments [17].
- The ICT-MMMO consists of online social review videos annotated at the video level for the sentiment [25].
- IEMO-CAP consists of 151 videos of recorded dialogues, with 2 speakers per session for a total of 302 videos across the dataset. Each segment is annotated for the presence of 9 emotions (angry, excited, fearful, sad, surprised, frustrated, happy, disappointed, and neutral) as well as valence, arousal, and dominance [4].

Arabic Standard Datasets

- ASTD (Arabic Sentiment Tweets dataset) which consists of about 10,000 tweets that are classified as objective, subjective positive, subjective negative, and subjective mixed. It covers a wide range of topics and provides a valuable resource for training and evaluating sentiment analysis models [15].
- HARD (The Hotel Arabic-Reviews dataset) is a data set specifically focused on sentiment analysis of hotel reviews written in Arabic. It is a collection of over 370,000 hotel reviews in modern standard Arabic and dialectal Arabic [7].
- ArabSign dataset which is a continuous ArSL dataset, consists of 9,335 samples performed by 6 signers. Where the annotation of the dataset was provided according to ArSL and Arabic language structures that can help in studying the linguistic characteristics of ArSL [14].
- AMMD (Arabic Multimodal dataset) which is compiled from YouTube videos, considering only video-blogging videos. The dataset attempts to include many different meta-information about the videos such as audio, visual gestures, transcript, and sentiment analysis annotation, all aligned with each other [2].

1.3 Related Work

Compared to non-Arabic languages, the availability of datasets for Arabic multimodal sentiment analysis is limited. While there may be datasets for unimodal sentiment analysis in Arabic, comprehensive datasets covering text, images, audio, and video are scarce. Some Arabic sentiment datasets have been collected such as:

- ASTD (Arabic Sentiment Tweets dataset), an Arabic social sentiment analysis dataset gathered from Twitter [15]. It comprises approximately 10,000 tweets that are classified as objective, subjective positive, subjective negative, and neutral sentiment.
  Nabil, Aly, et al. present the properties and the statistics of the data set and run experiments using standard partitioning of the data set. Their experiments provide benchmark results for 4-way sentiment classification on the data set.

- HARD (Hotel Arabic-Reviews dataset), a collection of over 370,000 hotel reviews in modern standard Arabic and dialectal Arabic[7]. The data set is available in two forms: the unbalanced complete set and the balanced data set. Each record includes the review text in Arabic, the reviewer's rating on a scale of 1 to 5 stars, and other attributes about the hotel and reviewer. The full unbalanced HARD dataset includes all the reviews, while the balanced subset data set includes 94,052 reviews with roughly equal numbers of positive and negative reviews.
  The authors tested six popular classifiers for polarity and rating classification, including Logistic Regression (LGR), Passive Aggressive (PAG), Support Vector Machine (SVM), Perception (PRN), Random Forest (RFT), and AdaBoost (ABT), and found that Logistic regression and SVM classifiers have achieved the best ac-curacies ranging from 94% to 97% for polarity classification, and 72% to 75% for rating classification. Additionally, the authors used a constructed lexicon to confirm the effectiveness of HARD and achieved an accuracy of 89% for polarity classification.

However, because of the large number of smartphones that eased the usage of audio-visual communication among individuals, unimodal sentiment analysis has become of little use and motivated researchers to move on to the next level which is Multimodal sentiment analysis.
Previous work on Multimodal Sentiment Analysis shows that it is a powerful tool for analyzing sentiment, as effective cues like facial and vocal expressions can reveal underlying emotions. By analyzing these cues alongside textual analysis, a more thorough comprehension of sentiment can be achieved.

- Poria et al. had introduced in their paper [18] a framework for Multimodal sentiment analysis that encompasses relevant features for both textual and visual data, as well as a simple technique for combining features from different modalities such as audio, visual, and text data. The proposed multimodal system in achieved an accuracy of nearly 80%, outperforming all of the previous art systems at that time by more than 20%.

- Soleymani et al.[20] presented a broad explanation of the ideas and aims of Multimodal sentiment analysis demonstrating that it is an effective method for utilizing different channels of information to analyze sentiment, outperforming the capabilities of Unimodal methods, which gives the potential to enhance other tools that depend on Unimodal sentiment analysis.

- Zadeh et al.[27] presented the CMU MOSEI dataset, abbreviated from "Multimodal Opinion Sentiment and Emotion Intensity," which is among the largest and most extensive multimodal datasets for sentiment analysis and emotion recognition. It includes over 23,000 annotated sentences from more than 1,000 online speakers covering 250 different topics. The authors investigated how modalities interact in sentiment analysis and emotion recognition, utilizing a new interpretable fusion mechanism called the Dynamic

Fusion Graph (DFG). By training the DFG in the Memory Fusion Network pipeline, DFG demonstrated excellent performance in sentiment analysis and competitive performance in emotion recognition.

- Abdulrahman Alqarafi et al. presented in 2017 a bridge to Multimodal sentiment analysis under the name (Towards Arabic Multimodal Sentiment Analysis) which was a novel Arabic multimodal dataset is presented and validated using a state-of-the-art support vector machine (SVM) based classification method at that time [2]. The dataset was built and validated using a state-of-the-art SVM-based classification method, with the aim of detecting the polarity from different models for the Arabic language. However, it contains only textual and video models.

- Multimodal sentiment analysis has played an important role in facilitating the work of several humanities and scientific fields, such as the hearing loss problem. In this context, Luqman presented in his paper "ArabSign: A Multimodality Dataset and Benchmark for Continuous Arabic Sign Language Recognition "[14], a dataset that consists of 9,335 samples performed by 6 signers. Where the annotation of the dataset was provided according to ArSL [1] and Arabic language structures that can help in studying the linguistic characteristics of ArSL. The dataset was acquired using Kinect V2 [2]. All samples are available in three formats: color, depth, and commonality. Also, Spatial features were extracted from sentence frames using two pre-trained models introduced in the proposed models. And finally, they evaluated the models on the proposed dataset in the signer-dependent and independent modes.

1.4 Summary and Conclusion

In this chapter, we began by introducing some preliminaries and definitions pertaining to Sentiment Analysis and its variants.
Following that, we delve into reviewing both Sentiment Analysis Approaches and datasets, with a specific focus on Arabic Sentiment Analysis datasets. The reviews summarized in Table 1.1 highlight several key observations:

- The scarcity of work considering the Arabic language.
- Especially, we can consider that Arabic Multimodal Sentiment Analysis is at its infantile age.
- When treated, modalities didn't consider the inherent semantics.

In order to investigate some of these lacks, in the next chapters we present our contribution that leverage Arabic Sentiment Analysis by both inherent semantic and multimodal.

[1] ArSL: Arabic Sign Language

[2] Kinect V2: A motion-sensing input device developed by Microsoft for gaming and other applications. It uses advanced depth-sensing technology to capture 3D information and track human movement, enabling various interactive experiences.

References	Language	dataset	Modality	Size	Source	Annotation	Pre-processing	Semantic
DT Vo et al. [15]	Arabic	ASTD	T	10,000 tweets	Twitter	Manual annotation using AMT[a], assigning sentiment labels to each tweet in the dataset. These labels indicate whether the sentiment expressed in the tweet is positive, negative, or neutral.	Removed noise, tokenized text, removed stopwords, normalized through stemming/lemmatization, handled negation, and addressed class imbalance, resulting in a clean, standardized dataset for accurate sentiment analysis.	Understanding the sentiment or emotions conveyed by users in their tweets, such as positive, negative, or neutral sentiment
A Elnagar et al. [7]	Arabic	HARD	T	370,000 reviews	Booking.com	Human annotators reviewed each hotel review and assigned sentiment labels (positive, negative, or neutral) based on the expressed opinions and emotions in the text	Removed noise (URLs, special characters), tokenized text into words, and applied normalization techniques (stemming/lemmatization) for sentiment analysis readiness.	-
H Luqman et al. [14]	Arabic	ArabSign	T/V	9,335 samples	-	It was annotated by experts who accurately labeled continuous Arabic Sign Language gestures using precise symbols or descriptions.	data cleaning, alignment, and normalization for accurate and robust continuous Arabic Sign Language recognition.	
A Al-Qarafi et al. [2]	Arabic	AMMD	T/V	830 segments	YouTube	categorizing transcribed sentences into subjective or objective utterances, and further annotating subjective utterances as positive or negative based on expressed opinions.	manual transcription of the videos in two levels, the initial transcription and the second transcriber reviewing and evaluating	-
Amir Zadeh et al. [26]	English	CMU-MOSI	T/A/V	2,199 segments	YouTube	assigning sentiment intensity and subjectivity labels to the opinion videos, capturing the range of emotions expressed by the speakers.	cleaning and transforming raw data, such as text transcriptions or multimodal features, into a suitable format for sentiment analysis and subjectivity analysis in online opinion videos.	It facilitates research into the semantic nuances of sentiment expression through textual, audio, and visual features.
Amir Zadeh et al. [27]	English	CMU-MOSEIE	T/A/V	23,453 utterances	YouTube	Manual annotation, Trained annotators carefully label the dataset with sentiment based on their subjective judgments.	Cleaning and transforming the multimodal data, including text, audio, and visual features by removing noise, normalizing, and ensuring consistency.	-
C Busso et al. [4]	English	IEMOCAP	T/A/V	10,039 record	-	Manual annotation, annotated with emotion labels, linguistic transcriptions, and other features.	-	

Table 1.1: Sentiment Analysis datasets

[a]Amazon Mechanical Turk

Chapter 2

Arabic Multimodal Dataset

In this chapter, we delve into the methodology that we used to construct our data set, which is inspired basically by the work of Zadeh et al [27]. Firstly, we discuss the process of data collection, presenting the different sources from which we collected our videos. We address also the various criteria we followed in detail to ensure the quality and relevance of the collected videos. Next, we explore the annotation process, discussing its different stages in depth. We highlight the tool used for annotation and explain how it facilitated the process. Following that, we define the various features of each modality, emphasizing the tools employed for their extraction. Finally, we discuss the alignment process, which involves synchronizing the different modalities. This step ensures the coherence and integration of the various data components.

2.1 Dataset Construction

Let us recall that our main objective in this work is twofold. Design a pipeline that facilitates the construction of an Arabic dataset that mainly enhances i) semantic understanding through the utilization of transformers and ii) the usage of multimodalities.

Our NLP task revolves around building a dataset. Figure 2.1 illustrates our designed pipe that is split into the following five phases:

1. **Collection:** in which we collect videos that align with our task goal.
2. **Segmentation:** in this step is dedicated to the chunking of the collected videos by capturing only the subjective parts and considering the rest of the parts as objective segments. Whose latter are also deployed to detect whether the opinion expressed is subjective or objective.
3. **Annotation:** the most complex phase where we annotate the subjective segments for emotion and sentiment recognition.
4. **Features Extraction:** In this step, we extract audio, text, and visual features
5. **Alignment:** alignment of Multimodal data, such as text, audio, and video

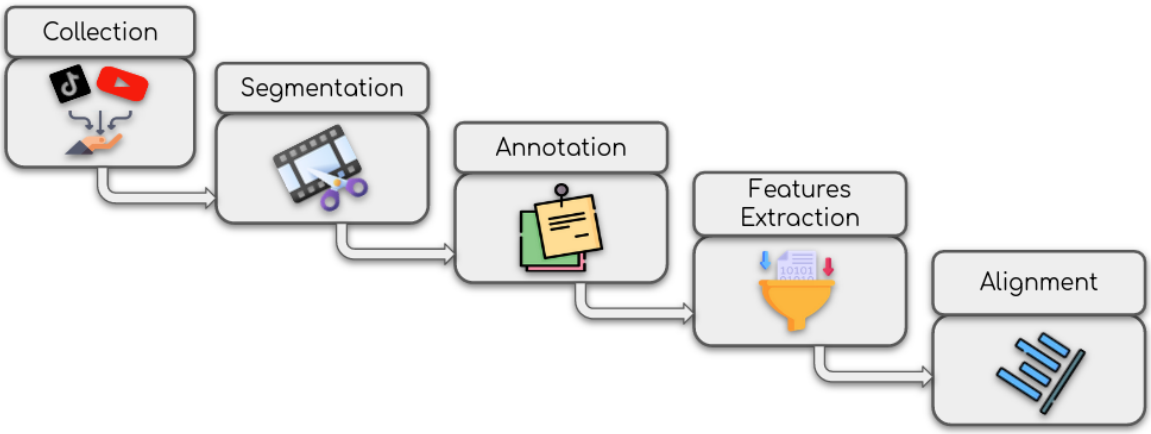


Figure 2.1: Five-steps pipeline.

2.1.1 Data collection

Since the aim of our work is multimodal sentiment analysis, social media sites were the most valid platforms for collecting videos, as they represent a unique opportunity to obtain large amounts of data from various speakers and topics as users often post their opinions in the form of videos for free.
In order to facilitate the search process we used YouTube API features to extract YouTube channel links based on specific topics such as Arabic, reviews, opinions, blogs, etc.
The collection process was carried out under many criteria, in order to reach the best and correct results. As these criteria are as follows

- The point on which all this work was built is the Arabic language, so it was important that the language used by the speaker in each passage be an original Arabic language or at least a dialect close to it.
- Another important condition is that the video must feature many meta information such as audio, and visual gestures.
- It is also required to ensure that each video or image contains only one face in order to avoid any ambiguity in the analysis and interpretation.
- Moreover, we tried to achieve another important criterion, which is diversity in topics and included the largest number of headlines such as politics, sports, entertainment, etc.
- With the need to respect that the voice and facial features of the speaker are clear and that there is no overlapping of the voices. Adding to that we tried to ensure the presence of both genders as speakers.

After adhering to all these criteria, about 72 videos were collected from two different platforms, 65 videos from YouTube and 7 from TikTok in the period between 10/03/2023 to 26/03/2023, respecting the condition that the video copyright license allows for academic usage in the public domain. Figure 2.2 displays a graph depicting the lengths in seconds of 72 videos.

The average length of the videos is about seven minutes. As mentioned previously, there should have been a diversity of topics raised in the videos, so we made sure that our research touches all permissible topics such as political, social, sports, cultural, and so on. Noting that after we selected the videos that were largely in line with the previously mentioned criteria, we again carried out the scrapping process, but this time it was about the contents of the video and the different topics that each one touches on.
It was also important that there was a difference in the level of the speakers themselves, as they totaled 66 speakers of both genders, 10 females and 56 males.

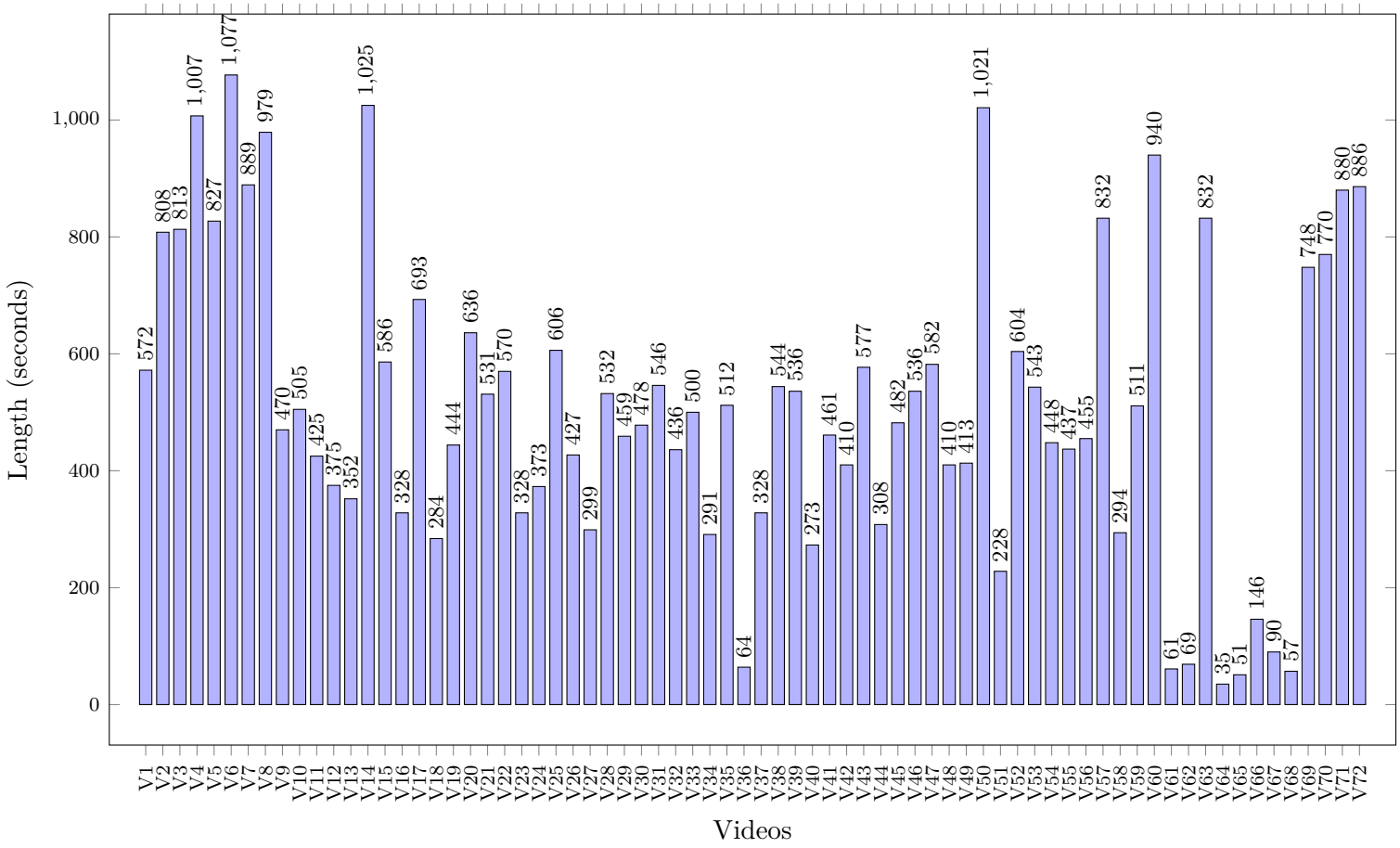


Figure 2.2: Length Of Videos Extracted by second

2.1.2 Annotation and Segmentation

We conducted our segmentation and annotation task manually using the Lab Home Multimodal Dataset Multitask Annotation Tool (MDMAT) [9]. It is an open-source software designed for segmentating and annotation temporal segments in video clips. Figure 2.3 illustrates MDMAT tool. The interface of the tool is divided vertically into three sections:

1. The right section: It consists of several components. It includes an uploader button, which allows users to upload video file. It contains also segmentation buttons that enable users to specify the start time and end time for each segment. Moreover, the right section also houses annotation buttons, such as sentiment scale, easy/hard, emotions, and body gestures.

2. The middle section: it displays the uploaded video clip along with its audio part, and it also visually represents the generated segments by coloring its corresponding portion in the audio section.

3. The left section: it provides information about the segments, including their timing and annotations. It also includes an import button that allows users who did not perform the segmentation task to import JSON files containing segmentation information from other users. This feature facilitates the unification of the number of segments and timing, ensuring consistency across users and enabling a collaborative approach to the annotation process.

The annotation team consists of three members, each member annotated all the collected videos on his own without any external consultation or interference. This means that the annotation task is passed on to three members.
In order to initiate the annotation process, we loaded our video file onto the MDMAT tool. Subsequently, we segmented the videos by defining the start and end times for each segment within the video file. It was agreed between the members of the annotation team on how to divide the video and annotate the segment, as defined in A guideline, as there must be a personal opinion and expression of feelings in the segment that will be annotated, which will be considered subjective opinion. The remaining segments will be considered objective opinion.
Additionally, we made sure to avoid allowing our individual opinions to influence one another during the annotation process. The MDMAT tool offers an import feature that facilitated collaboration and sharing of the annotation process among team members. This feature streamlined the workflow, allowing for efficient coordination and exchange of segment data without sharing annotation data. Each segment was assigned a unique ID, which simplifies the annotation process. By selecting the segment ID, we proceeded with annotating the segments. When each member finishes annotating a particular video, he/she uploads the JSON file from MDMAT containing his own annotation data attached to his name and sends it to the other members. Subsequently, each member imports the segments and made their own annotations, repeating the process until the JSON file contained annotations from all members. The JSON file included the number and duration of segments for each video, as well as the name and annotations of each member. After all the members have finished annotating the videos, the JSON files are collected to initiate the annotation merging process, aiming to obtain a unified annotation for each video. Annotating segments of the videos was the most challenging task in our mission due to the Multitask Annotation approach, which involved dividing the annotation process into four sections:

- The first section focused on determining the intensity of sentiment, where we utilized the sentiment scale [-3, -2, -1, 0, 1, 2, 3] to represent highly negative, moderately negative, slightly negative, neutral, highly positive, moderately positive, and slightly positive sentiments, respectively. The values of the scale represents the intensity of sentiment of the speaker

- The second section centered around identifying the speaker's emotions within each segment. The emotions considered in this section include happiness, pride, excitement, gladness, interest, sadness, anger, afraid, disgust, sorry, surprise, funny, nervous, loving, hating, and boredom. Determining the intensity of sentiments relies on identifying the corresponding emotions, and vice versa. It is important to note that a negative intensity of sentiment cannot be paired with positive emotion, and vice versa. Additionally, it was

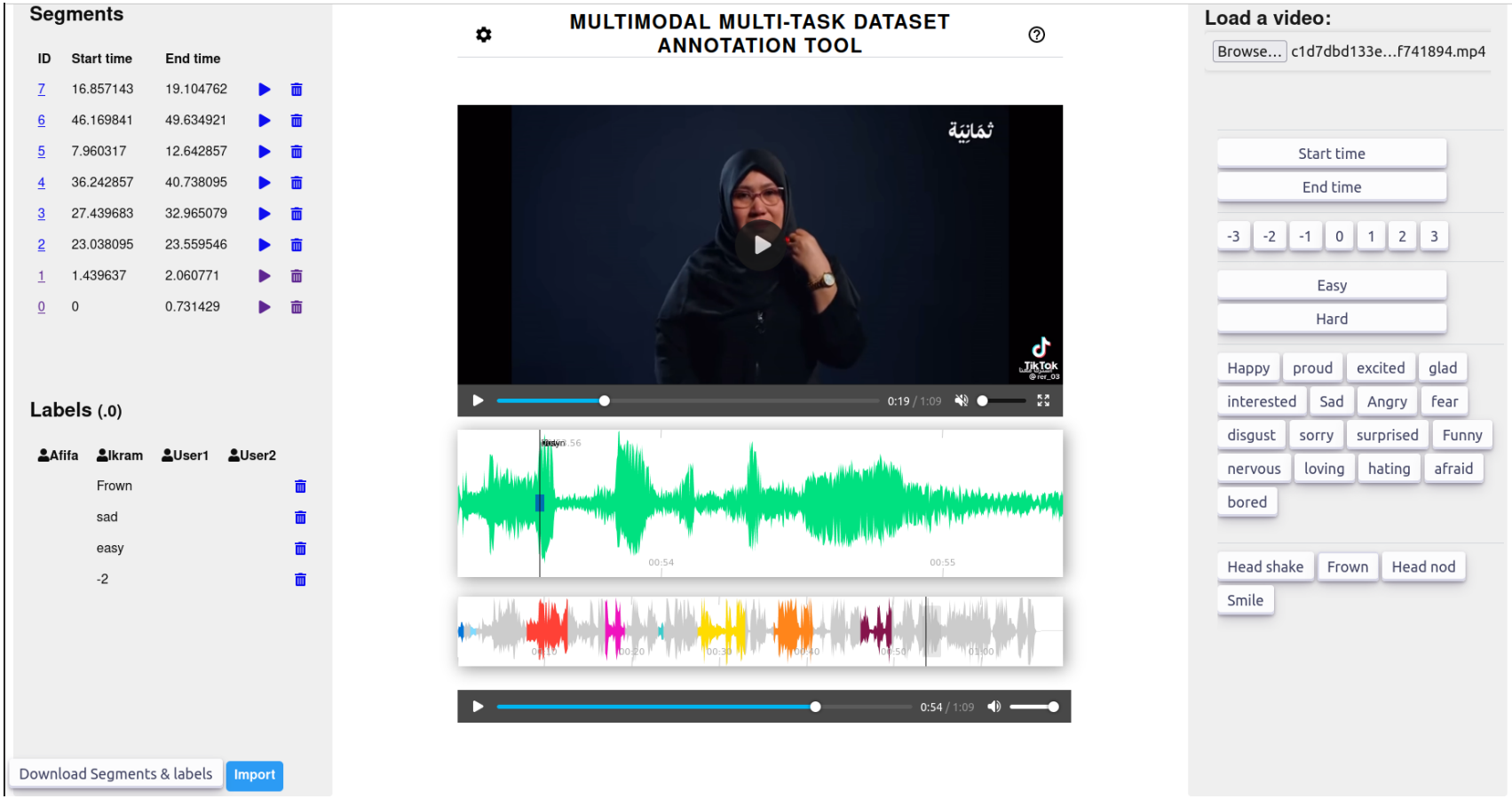


Figure 2.3: The Multimodal Dataset Multitask Annotation Tool interface

agreed that in cases of confusion or difficulty in determining the speaker's emotion, the happy emotion should be chosen when a positive intensity of sentiment is present, while the sad emotion should be chosen when a negative intensity of sentiment is observed.

- The third section revolves around choosing one of the following options: "hard" or "easy". Where the option "easy" means that the speaker's emotion and sentiment intensity were easily identified from speech only, while the option "hard" means that we needed the facial and acoustic features to determine it.

- The fourth section is the simplest section, it focuses on identifying and labeling specific gestures and facial expressions exhibited by the speaker during the video segment, as it consists of four options which are: Head shake, Head nod, smile, and frown. where one of these options is chosen according to the movements and expressions made by the speaker. The annotation task was the most difficult task in our mission and the most time-consuming task because it is difficult to determine the intensity of human sentiment and emotions.

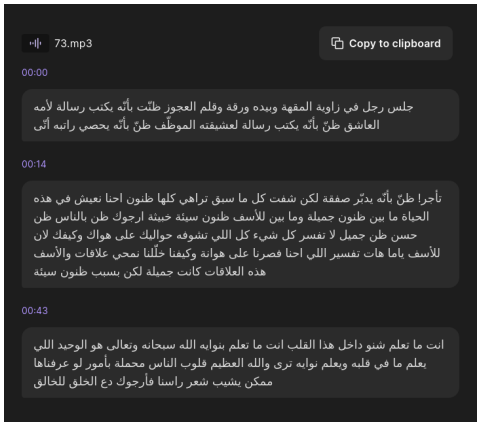
2.1.3 Features extraction and data representation

Text

To transcribe the Arabic speech in our videos according to the text modality, we initially extracted the audio files from the video files using the ffmpeg open-source command-line tool. Subsequently, we utilized the Almufaragh App [1] to extract the Arabic transcriptions from the audio files. Almufaragh is a free Windows application specifically designed for Arabic speech recognition, as shown in Figure 2.4a. We inputted our audio files into the app, which generated SRT and text files as output. The files that were created by Almufaragh App were not satisfactory for us, so we used a web AI-based application, AI Transcription[2], which is a free web application developed by Riverside to extract the Arabic transcriptions. Figure 2.4b shows the AI Transcription tool interface in action. AI transcriptions automatically transcribe the audio and video recordings with the AI-powered technology. This advanced technology is specifically designed to efficiently process a wide range of file formats, including popular formats like MP4, MP3, WAV, and MOV. Moreover, AI transcriptions offers extensive language support, encompassing over 100 languages, including the Arabic language. This enables users to accurately transcribe content in various languages. The utilization of this technology in our task facilitated the transcription process, allowing for efficient and accurate conversion of audio and video recordings into written text. The final step in the transcription phase was to verify the automatically extracted transcripts, in which we manually validated the written text and timings for all the files.
After verifying the validity of the texts extracted from the videos we moved on to the next step which is the word embedding technique. Word embedding is a real-valued vector representation of words by embedding both semantic and syntactic meanings obtained from an unlabeled large corpus. It is a powerful tool widely used in modern natural language processing (NLP) tasks, including semantic analysis, information retrieval, dependency parsing, question answering, and machine translation [24]. In order to achieve this phase, we chose Arabert as a tool. Arabert is a pre-trained Arabic language transformer model based on BERT [1]. we simply run the program which automatically returns the results in two different formats, tsv files and txt files.

[1] `https://almufaragh.com/`
[2] `https://riverside.fm/transcription`

(a) Almufaragh Application interfaces



(b) The AI Transcription tool interface

Figure 2.4: The transcription tools interface

Audio

Acoustic features play an important role in capturing and analyzing the acoustic component of the input data, as they provide valuable information about the speaker's emotional state, tone of voice, and tone of voice, which can greatly contribute to the overall sentiment analysis task. In our study, we first extracted all audio from previously collected videos then we used the openSMILE feature extraction toolkit, which unites feature extraction algorithms from the speech processing and the Music Information Retrieval communities[**?**], where There were several acoustic features that were extracted, including the following:

- Mel-frequency cepstral coefficients (MFCCs): are a set of features commonly used in audio and speech processing to capture the spectral characteristics of a sound signal.
- Voice quality feature: It refers to the overall sound and timbre of the voice, including factors such as tone, clarity, resonance, and texture.
- Emotion-related feature refers to a measurable characteristic or attribute that captures or represents aspects of an individual's emotional state.
- Prosody feature: refers to how we say things rather than what we say. It involves the modulation of our voice, the way we stress certain words or syllables, and the patterns of rising and falling pitch that convey meaning, emotion, and intention in communication.
- These features provide valuable information about the shape and characteristics of the vocal tract during speech production.

- Spectral features: capture information about how the energy or power is distributed across different frequencies in a signal.
- Timing features: provide information about the timing and duration of specific events or actions.

Visual

Visual features are one of the most important ways in which people express their feelings. These features are divided into two parts. The first part is body gestures, which focus on the physical characteristics of the speaker's body, such as shaking the head, crossing or spacing the arms, and gestures. The second part is the facial features, which are Facial movements and expressions, such as smiling, frowning, widening eyes, and other expressions that an individual makes to express his feelings. Since humans rely on facial features a lot to express their feelings in a clearer way, we also relied on them in our work, in addition to that facial features can be extracted and captured easily compared to body gestures. We relied on OpenFace tool for the visual feature extraction task, as shown in Figure 2.5, which is an open-source tool designed by MultiComp Lab for researchers in the field of computer vision [3]. The main features that we

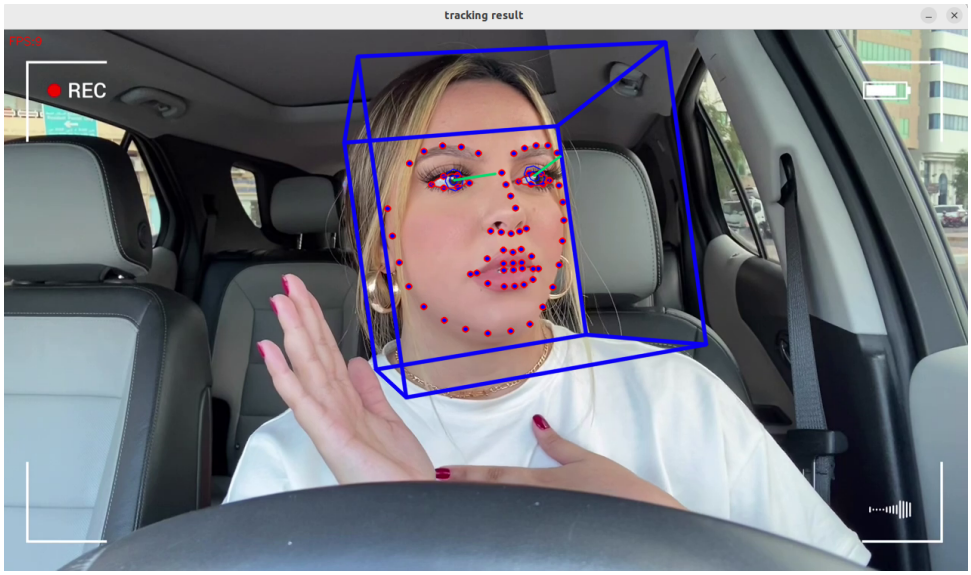


Figure 2.5: OpenFace Tool Interface

rely on and extract through OpenFace are:

- Head pose: it involves the position and orientation of the head in three dimensions, specifically capturing its pitch if it tilts up or down, its yaws if it rotations left or right, and its rolls if it tilts side to side.
- Action Unit: includes facial movements and expressions such as Inner or Outer brow raiser, Chin raiser, Nose wrinkler, and eye closure.
- Facial landmark: this includes capturing specific points on a face and tracking their movement that is used to identify facial features and landmark shape variations.
- Eye gaze: it is the detection of eye-region landmarks. This includes eyelids, iris, and the pupil.

2.1.4 Alignment

To synchronize the text files generated by Arabert with our audio files, we cleaned first the text files from unnecessary data by a simple command, then we used Aeneas. Aeneas is an open-source forced aligner software used to automatically generate a synchronization map between text fragments and audio, without any human involvement. It defines the time interval in the audio for each word or text fragment, resulting in a time-stamped representation of speech.
After extracting features from each modality and obtaining the time-stamped representation of speech through the forced alignment process, our final step involved aligning data from multiple modalities. To facilitate this, we utilized the CMU Multimodal Data SDK [26], a Python SDK developed by the MultiComp Lab. The CMU Multimodal Data SDK is specifically designed to simplify the loading and alignment of multimodal data. By taking advantage of the CMU Multimodal Data SDK features, we aligned and structured the data from various modalities into a dictionary format. This dictionary format consisted of multiple computational sequences, each representing a specific modality.

Figure 2.6 summarizes the used software in each of the five steps described above.

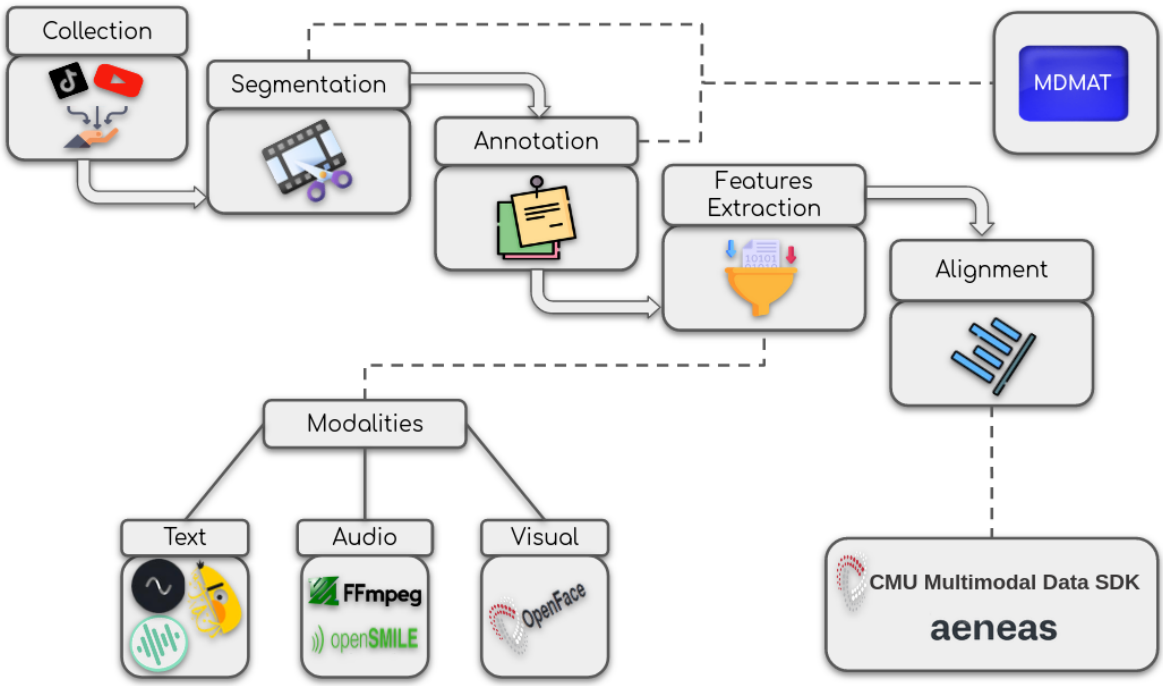


Figure 2.6: The Used Tools/Softwares for the Pipe' Implementation .

2.2 Summary and Conclusion

In this chapter, we have described in detail the adopted Methodology to built our AMMD dataset dedicated to the development and validation purpose of some NLP task's Models based on AI. Specially, the tasks are Multimodal sentiment analysis, Multimodal Subjectivity detection and Multimodal emotion recognition.

In fact, from social media we have scrapped a set of chosen videos that we have processed in order to get our dataset. Following some recent dataset building engineering, we have deployed nine recent and innovative tools. Intentionally, the annotations are done manually and validated using Agreement Annotator Method to be more effective.

Chapter 3

Dataset Assessment

In this chapter, we aim to assess the quality of the combined AMMD dataset. We will begin by providing a detailed and comprehensive description of the dataset while highlighting its key features and characteristics. In a second section, we evaluate the dataset through the main NLP task multi-modality Sentiment Analysis. A justification for the chosen deep learning model is provided. Finally, we will present the experimental results.

3.1 Description of the AMMD Dataset

The Arabic MultiModal dataset (AMMD) has been successfully created through our data-building pipeline, resulting in a resource for conducting multimodal sentiment analysis specifically tailored for the Arabic language.

In this section, we will delve into the dataset's statistics, quality, and the diverse topics it covers, presenting a comprehensive overview of its attributes. Table 3.1 summarizes some key details on the AMMD dataset.

- **Size**: The AMMD dataset with a size of 80,2MB, it originally consists of 72 video extractions that were collected during the data collection process, resulting in a total video duration of 9 hours, 59 minutes, and 30 seconds. From which a total of 976 video segments were extracted.
  After thorough data preparation and annotation, we obtained a refined dataset comprising 488 subjective segments extracted from the collected videos. These segments represent distinct portions of the videos that capture subjective expressions and emotions.
  Figure 3.1 visually demonstrates the distribution of video segments within the dataset, showcasing the varying number of segments across different videos.

- **Segment Statistics**: The AMMD dataset comprises a total of 488 subjective segments, extracted from the collected videos.
  These segments have an average length of 2.82 seconds. Among these segments, there are 238 positive segments, 223 negative segments, and 2 neutral segments.
  The distribution of the AMMD segments in terms of sentiment is visualized in Figure 3.4. The chart showcases the relative proportions of positive, negative, and neutral segments, providing a visual representation of the sentiment composition within the dataset. One

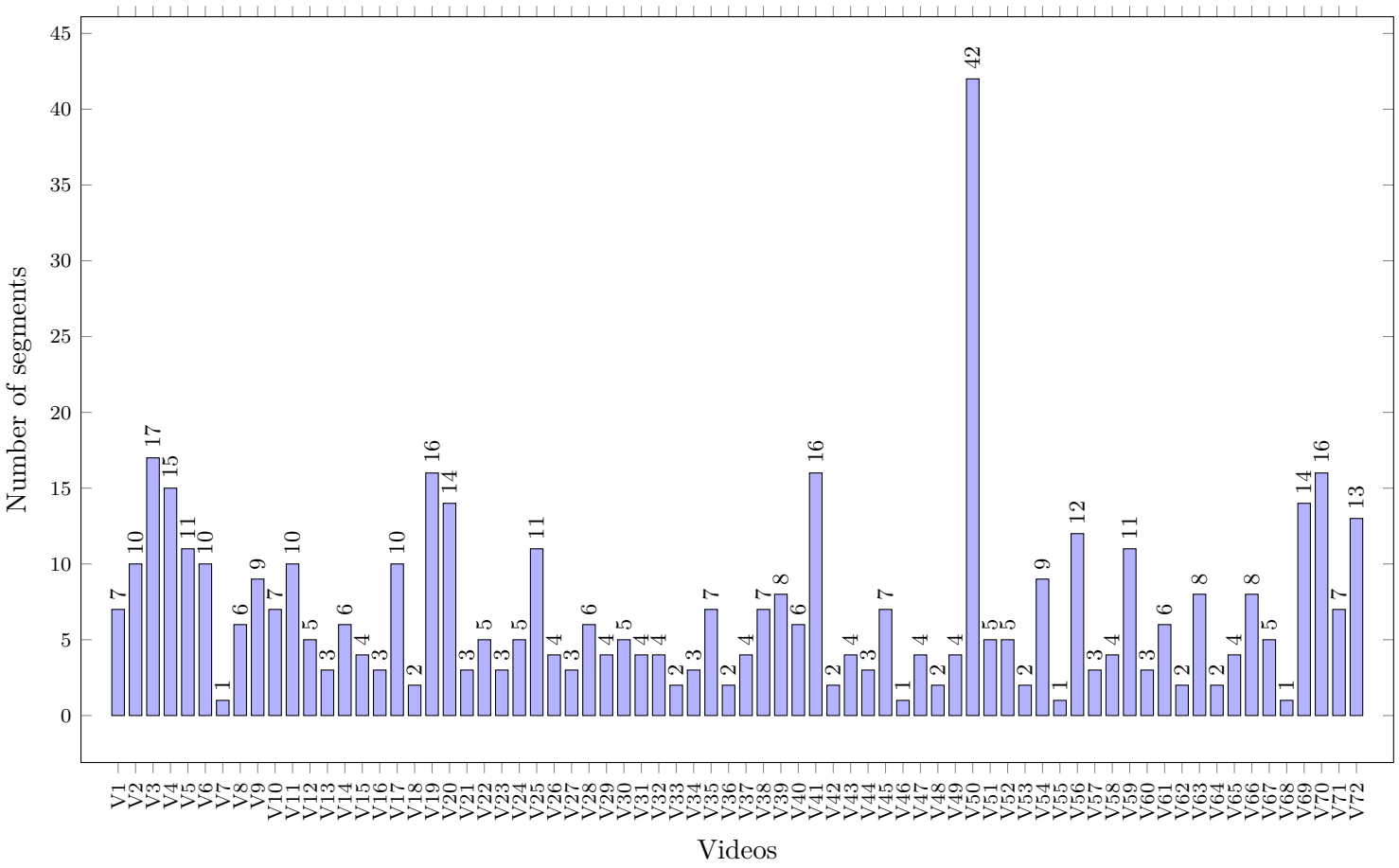


Figure 3.1: Number of Segments extracted from processed videos

can observe that neutral segments need to be enlarged to performed a more balanced dataset.

- **Speaker Statistics**: The AMMD dataset consists of segments from a diverse group of 50 speakers. Among these speakers, there are 9 female speakers and 41 male speakers. The inclusion of both male and female speakers adds to the richness and variety of the dataset. The distribution of speakers by gender can be visualized in the pie chart Figure 3.2, which illustrates the relative proportions of female and male speakers within the dataset.

- **Variety of Topics**: The AMMD dataset encompasses a wide range of topics, providing a diverse collection of videos for sentiment analysis. This dataset includes segments covering 11 different topics, ensuring the inclusion of various subjects. The topics covered in this study may include politics, sports, entertainment, culture, social issues, and more.
  The distribution of topics within the AMMD dataset can be visualized in 3.3, illustrating the relative proportions of different topics present in the dataset.

- **Number of Words**: The AMMD dataset comprises a total of 66, 334 different words across the collected videos. These words encompass the verbal expressions within the segments.

- **Multimodal Features**: The AMMD dataset incorporates multimodal features, including textual, visual, and acoustic information. The textual features are represented by embedding dimensions of 768, capturing the semantic and contextual information within the

segments. The visual features comprise dimensions of 714, encoding visual cues such as facial expressions and gestures. Additionally, the acoustic features encompass dimensions of 54, capturing the acoustic properties of the audio, such as pitch, intensity, and voice quality.

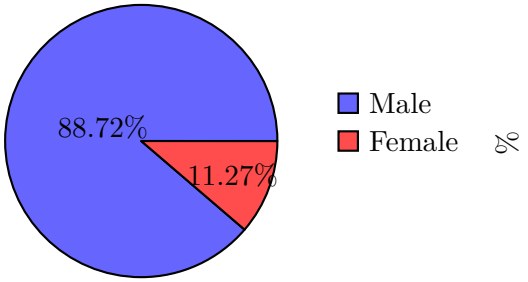


Figure 3.2: Gender Statistics

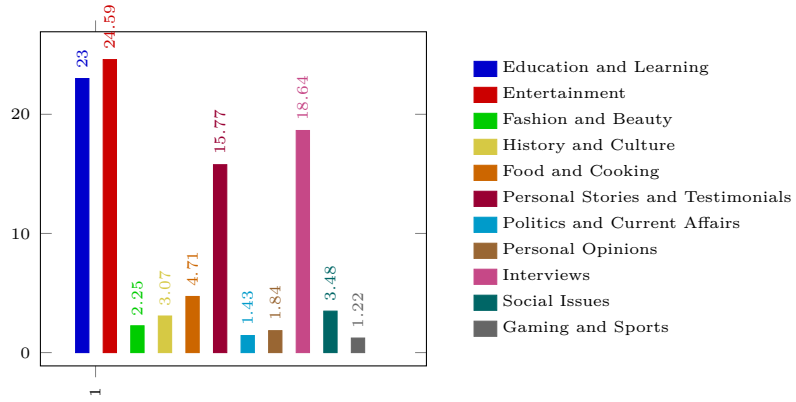


Figure 3.3: Topics Statistics

3.2 Model Selection

Let us recall that the AMMD dataset we built in order to train NLP multimodal models mainly incorporates semantic information and target Sentiment Analysis, Subjectivity and Emotion Recognition applications.

The inherent semantic is mainly captured through transformers and multimodalities usage which we relied on for assessing our dataset.

In order to assess our dataset, we have chosen a state-of-the-art architecture that fuses modalities early, known as the Multimodal Transformer (MulT) [22]. MulT is a transformer-based architecture that efficiently handles and integrates different modalities. Figure 3.5 illustrates the architecture of the Multimodal Transformer. MulT leverages the power of the attention mechanism that enable capturing relationships and dependencies within multimodal data, further enhancing its ability to understand the semantic information present in the dataset.

This Model has proved its efficiency for many Multi-modality-based NLP tasks. More details are provided in [22].

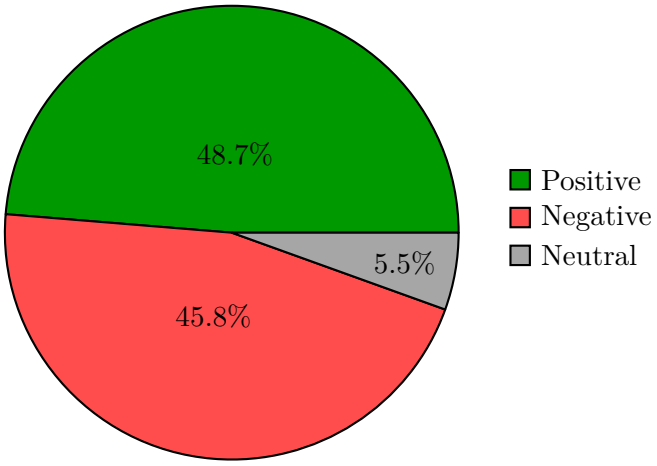


Figure 3.4: Distribution of the AMMD segments in terms of Sentiment

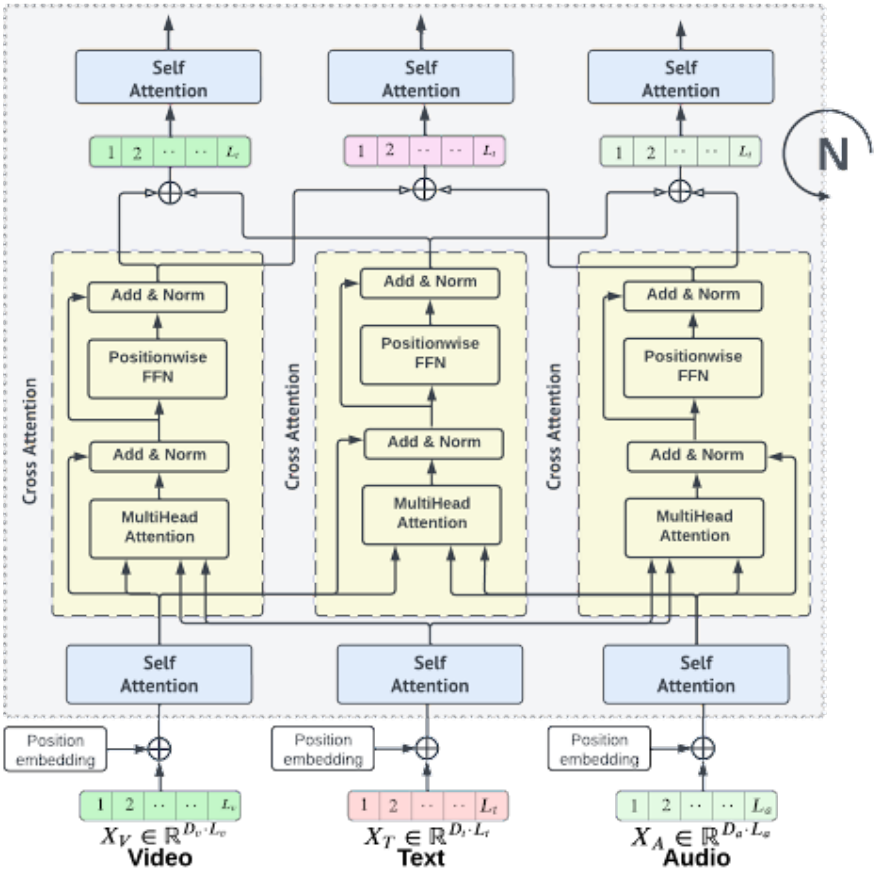


Figure 3.5: Multimodal Transformer Architecture

Dataset Description	
Size	80,2 MB
Number of videos	72
Total number of segments	976
Total number of subjective segments	488
Number of words	66334
Total videos time	09h :59min : 30s
Average length of segments	2.82 seconds
Number of positive segments	238
Number of negative segments	223
Number of neutral segments	27
Number of speakers	50
Number of female speakers	9
Number of male speakers	41
Text embedding dimensions	768
Visual Feature dimensions	714
Acoustic Feature	54

Table 3.1: Dataset Statistics and Description

Transformer Components

The Transformer architecture consists of sundry key components that work with each other to process and transform input sequences.

Input Layer

The input layer in a neural network architecture serves as the initial stage where the raw input data is processed and prepared for further processing. In the context of the MulT architecture, the input layer which called the input embedding layer is responsible for transforming the raw input data into embedded representations. The positional encoding is applied then to

the embedded representations to provide information about the position or order of the input elements within the sequence, which result a set of embedded representations with positional information that are then passed as input to the self-attention layer [22].
The purpose of positional encoding is to inject positional information into the input embeddings, as the transformer model does not inherently have any notion of sequence order [22].

Self-Attention And Cross-Attention

Attention is a fundamental mechanism in neural networks that allows models to selectively focus on relevant information while processing input data. It enables each element of the input sequence X_i, where "i" represents the modality among text, video,audio ,to attend to all the other elements. This process is referred to as the self-attention, it generates a new sequence $\widehat{X}_i$, where each element is weighted based on its relationship with other elements [22].
The calculation of self-attention starts by obtaining the Query (Q), Key (K), and Value (V) vectors through the multiplication of the embedded input sequence by their respective weight matrices. Once the Query (Q), Key (K), and Value (V) vectors are generated, the relevance between Q and K is measured, resulting in attention scores. These attention scores are calculated by taking the dot product of Q and K and dividing it by the square root of the dimension of the key vectors (dk), denoted as $\sqrt{d_k}$. This scaling factor ensures that the attention scores are appropriately scaled. The Softmax function is then applied to these scores to obtain attention weights. These weights are used to compute a weighted sum of the Value vectors, producing the final output representation. The equations 3.1 and 3.2 below describe the generation of $\widehat{X}_i$ [22].

$$\begin{aligned} &\text{Query (Q) vectors:} \quad Q = X_i \cdot W_q \\ &\text{Key (K) vectors:} \quad K = X_i \cdot W_k \\ &\text{Value (V) vectors:} \quad V = X_i \cdot W_v \end{aligned} \tag{3.1}$$

$$\widehat{X}_i = \text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) V_j \tag{3.2}$$

The Cross Attention is indeed an attention mechanism that operates between different sequences or modalities, and it typically uses different sizes for the query (Q), key (K), and value (V) vectors compared to self-attention[22].
In self-attention, the query, key, and value vectors are usually of the same size and are derived from the same input sequence. Each token in the input sequence generates its own query, key, and value vectors, allowing it to attend to other tokens within the same sequence. However, in cross-attention, the query vectors (Q) are often generated from one sequence or modality, representing the tokens that need to attend to information from another sequence or modality. The key (K) and value (V) vectors, on the other hand, are derived from the other sequence or modality, providing the information to be attended to. In this case, the dimensions of the Q, K, and V vectors can differ, allowing the model to capture dependencies and relationships between different modalities or sequences. [22].

The Multi head Attention

The Multi-head Attention involves multiple sets of self-attention calculations, where each head performs a separate self-attention calculation with its own set of weight matrices. Equation 3.3 describe the output multi-head attention of Mult model . Each set is called a "head". By employing multiple heads in the cross-attention mechanism, the model can capture different

aspects of the relationship between modalities. Each head can focus on specific interactions between modalities, allowing for a more comprehensive understanding of their dependencies. The outputs of the individual heads are then concatenated and linearly transformed using the output weight matrix [22].

$$\begin{aligned} \text{MultiHeadAttention}(X) &= \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W_O \\ \text{head}_k &= \text{Attention}(X_i W_{Qk}, X_i W_{Kk}, X_i W_{Vk}) \end{aligned} \tag{3.3}$$

The Normalization and Residual layer

The Normalization and Residual layer: Noramlization layers guarantee that the values within each layer of the trasnformer have consistent distribution, while Residual layers preserve the original input data and ensure that it flows through the network unchanged, or with minimal alteration. Both of these techniques are crucial for the constancy and efficiency of the Mult model [22].

The position-wise Feed-Forward Neural Networks

it applies a non-linear transformation to each element in the sequence independently. It helps catch complex patterns and adapt to different positions, meliorativing the model's performance in various tasks. The position-wise feed-forward neural network is represented by the equation 3.4, where Y represents the input vector, W_1 and W_2 are weight matrices, b_1 and b_2 are bias terms, and ReLU denotes the rectified linear unit activation function. [22].

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{3.4}$$

After normalizing the output of the position FFN networks, the resulting representations from each modality are typically concatenated together. This concatenation allows the model to capture the combined information from multiple modalities. The concatenated representations are then passed through additional layers. This process is repeated for multiple iterations, till the model enhances its understanding of the relationships and dependencies between all modalities.

3.3 Experimental Results

In order to assess the performance of the MulT model, we have relayed on a set of metrics and we have deployed the implementation [1] of the model due to Haouhat et al. [10].

Metrics

We used several metrics, including accuracy, F1 score, mean absolute error (MAE), and correlation coefficient. Let's review the various metrics used to evaluate performance or measure different aspects in a given context.

- Accuracy: is a commonly used metric to evaluate the performance of classification models [11]. It measures the proportion of correctly predicted segments out of the total segments in the dataset.

[1] `https://github.com/belgats/Arabic-Multimodal-Dataset/`

$$\text{Accuracy} = \left(\frac{\text{Number of Correctly Classified Segments}}{\text{Total Number of Segments}}\right) \times 100$$

- F1 score: This is a measure that combines precision which is the proportion of correctly predicted positive segments (true positives) out of all the segments predicted as positive, and recall which measures the proportion of correctly predicted positive segments out of all the actual positive segments in the dataset, into a single metric.

$$\text{F1 score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

- MAE (Mean Absolute Error): measures the average absolute difference between the predicted and actual values [5].

$$\text{MAE} = \frac{1}{n} \sum |\text{Predicted Value} - \text{Actual Value}|$$

- Correlation Coefficient: measures the strength and direction of the linear relationship between two variables [13].

$$\text{Correlation Coefficient} = \frac{\text{Covariance}(X, Y)}{\text{Standard Deviation}(X) \times \text{Standard Deviation}(Y)}$$

Results

The performance evaluation of our model is summarized in 3.2, which yielded the following results:

- Accuracy: The accuracy of 51.28% suggests that slightly more than half of the segments were correctly classified by the model.
- F1 score: The F1 score of 50.46% indicates that the model achieved a balance between precision and recall.
- MAE (Mean Absolute Error): The MAE of 1.298 suggests that, on average, the model's predictions differed by approximately 1.3 units from the true sentiment values.
- Correlation Coefficient: The correlation coefficient of 0.063 reveals a very weak positive linear relationship between the predicted and actual sentiment values.

Overall, the results obtained from running the MUlt transformer model on the AMMD dataset indicate a moderate performances compared to an average accuracy of 60% in multimodal sentiment analysis for other languages[22, 2].

One can argue that these moderate performances, for this first version of AMMD, are due to those factors:

- The accuracy of the annotation process that uses sentiment intensity. In fact positive sentiment can be intense in the range of +3, +2, or +1. The system considers that an annotation of +3 is different from +2 despite they are both positive.
- Arabert transformer is well suited for Modern Standard Arabic while most youtube' videos are more dialectal. That suggest using a Bert version that consider Arabic dialects semantics.

Metric	**Value**
Accuracy	0.512
F1 Score	0.505
MAE	1.298
Correlation Coefficient	0.063

Table 3.2: Performance Metrics

Conclusion and Perspectives

In this thesis, we have presented the AMMD dataset, a dataset that may represent an addition in Arabic multimodal research, particularly of sentiment analysis. The AMMD dataset offers a comprehensive collection of 488 annotated segments from over 50 online speakers, covering 11 different topics. Its construction involved a carefully executed methodology encompassing data collection, multi-label annotation, feature extraction, and data alignment. Each step is carefully executed to ensure the dataset's quality, reliability, and relevance to the field of Arabic multimodal research.

Before delving into the methodology, we provided an introduction highlighting the importance of Multimodal Sentiment Analysis Datasets, both in general and specifically in Arabic language processing.
We also included a discussion on various modalities and an overview of existing datasets in both Arabic and non-Arabic contexts. Additionally, we presented a summary of related work in the field of Multimodal Sentiment Analysis Datasets, which provided a broader understanding of research gaps.

In the end, The dataset's composition, size, and diversity were thoroughly described, and a carefully chosen model was employed for training, evaluation, and validation the dataset. The experimental results obtained from analyzing the AMMD dataset showcased its potential and practical applications in the field of natural language processing, particularly in sentiment analysis within the Arabic language domain.

In conclusion, in future iterations of the Arabic Multimodal Dataset for Sentiment Analysis (AMMD), additional perspectives can be integrated to further enrich its scope. Such as:

- Incorporating more dialects and languages: Including sentiment annotations from various Arabic dialects and languages to capture the nuances and variations in sentiment expression across different linguistic contexts.
- Expanding the dataset size and covering diverse topics: According to the time allowed to that task, the reached dataset size remains modest specially when we consider AI-based targeted models. For that, in ongoing work, efforts are being made to increase the dataset's volume by including a larger number of instances. This expansion aims to incorporate a wide range of topics, ensuring the dataset's diversity and enhancing its generalizability across different domains.
- Reducing gender bias: Ensuring a balanced representation of male and female speakers within the dataset to mitigate any potential gender-based differences in sentiment analysis results.

- Including additional annotation labels: Augmenting the dataset with annotations that go beyond facial expressions and encompass body and head gestures, enabling a more comprehensive analysis of multimodal cues and their impact on sentiment analysis.

By incorporating these perspectives into future iterations of the Arabic Multimodal Dataset for Sentiment Analysis (AMMD), it can become a more inclusive, representative, and comprehensive resource for studying sentiment in Arabic language across various dialects, topics, genders, and expressive modalities.

Bibliography

[1] Abdullah M Abu Nada, Eman Alajrami, Ahmed A Al-Saqqa, and Samy S Abu-Naser. Arabic text summarization using arabert model using extractive text summarization approach. 2020.

[2] Abdulrahman S Alqarafi, Ahsan Adeel, Mandar Gogate, Kia Dashitpour, Amir Hussain, and Tariq Durrani. Toward's arabic multi-modal sentiment analysis. In *Communications, Signal Processing, and Systems: Proceedings of the 2017 International Conference on Communications, Signal Processing, and Systems*, pages 2378–2386. Springer, 2019.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[5] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[6] Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: A survey of methods, trends and challenges. *ACM Computing Surveys*, 2023.

[7] Ashraf Elnagar, Yasmin S Khalifa, and Anas Einea. Hotel arabic-reviews dataset construction for sentiment analysis applications. *Intelligent natural language processing: Trends and applications*, pages 35–52, 2018.

[8] Giuseppe Fenza, Mariacristina Gallo, Vincenzo Loia, Francesco Orciuoli, and Enrique Herrera-Viedma. Data set quality in machine learning: consistency measure based on group decision making. *Applied Soft Computing*, 106:107366, 2021.

[9] Abdelhamid HAOUHAT. Mdmat : The multimodal dataset multitask annotation tool, 2023. Accessed on April 20, 2023.

[10] Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, and Hadda Cherroun. Towards arabic multimodal dataset for sentiment analysis. *arXiv preprint arXiv:2306.06322*, 2023.

[11] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[12] Intisar O Hussien and Yahia Hasan Jazyah. Multimodal sentiment analysis: A comparison study. *Journal of Computer Science*, 14(6):804–818, 2018.

[13] Yingbo Liu, Jiujun Cheng, Chendan Yan, Xiao Wu, and Fuzhen Chen. Research on the matthews correlation coefficients metrics of personalized recommendation algorithm evaluation. *International Journal of Hybrid Information Technology*, 8(1):163–172, 2015.

[14] Hamzah Luqman. Arabsign: A multi-modality dataset and benchmark for continuous arabic sign language recognition. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.

[15] Mahmoud Nabil, Mohamed Aly, and Amir Atiya. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2515–2519, 2015.

[16] Alvaro Ortigosa, José M Martín, and Rosa M Carro. Sentiment analysis in facebook and its application to e-learning. *Computers in human behavior*, 31:527–541, 2014.

[17] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, 2013.

[18] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[19] Anasua Sarkar, Yang Yang, and Mauno Vihinen. Variation benchmark datasets: update, criteria, quality and applications. *Database*, 2020, 2020.

[20] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[21] Iryna Sydorenko. What is a dataset in machine learning. *labelyourdata. com*, 2021.

[22] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[23] G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

[24] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8:e19, 2019.

[25] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.

[26] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[27] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

Appendix A

Annotation Guidelines

In order to uniform the segmentation and annotation, the annotators have to respect the following rules:

1. The segment must contain an opinion or an expression of feelings, whether through speech or facial features.
2. The length of the segment should not exceed 10 seconds.
3. The segment must contain sound language and clear and understandable Arabic words.
4. If it is difficult to determine the speaker's emotion, the degree of sentiment is considered. If it is negative, the emotion "sad" is chosen, and if it is positive, the emotion "happy" is chosen.
5. Priority is given to facial expressions such as "smile" and "frown", and if none of them are available, body gestures such as "head shaking" and "head nod" are considered.