



**Arabic Speech Sentiment Analysis Using Machine Learning: A
Case Study of COP28**

تحليل مشاعر الكلام العربي باستخدام التعليم الآلي: دراسة حالة كوب 28

by

SHEIKH ABDULAZIZ ALMUALLA

Dissertation submitted in partial fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

June 2024

Declaration

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the contents of this dissertation for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study, or education, and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my dissertation for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.



Signature of the student

Copyright and Information to Users

The author whose copyright is declared on the title page of the work has granted to The British University in Dubai the right to lend his/her research work to users of its library and as an open source, and to make partial or single copies for educational and research use. The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of the preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar, or the Dean only.

Copying for financial gain shall only be allowed with the author's express permission. Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

Abstract

In recent years, the UAE has played a pivotal role in advancing the global climate agenda by hosting significant events such as the COP28. COP28 served as a crucial platform for international dialogue and cooperation among nations to address climate change and accelerate efforts to mitigate its impacts. In an era characterized by rapid technological advancements, the development of Arabic speech recognition systems emerges as a crucial frontier in enhancing accessibility, efficiency, and usability across various domains. Despite significant advancements in speech recognition for languages like English, challenges persist in adapting these technologies effectively to accommodate the unique characteristics of Arabic. Within this context, exploring Arabic speech recognition within the framework of COP28 serves as a compelling case study. This research integrates speech recognition technologies at COP28 and holds the potential to streamline communication and enhance accessibility for Arabic-speaking delegates and stakeholders. Through a comprehensive investigation of various speech recognition models, including CNN, BI-LSTM, GRU, and hybrid architectures such as CNN-BI-LSTM and GRU-BI-LSTM, valuable insights can be gained into their performance and efficacy within the unique context of Arabic speech recognition. Analysing key metrics such as accuracy across different sentiment categories – positive, negative, and neutral – provides a nuanced understanding of each model's strengths and limitations. The hybrid GRU and BI-LSTM model takes the lead, showcasing outstanding performance with an accuracy rate of 94%. Close behind is the standalone GRU technique, achieving an accuracy of 93%. Subsequently, both the CNN-BI-LSTM and CNN models follow suit with accuracies of 91% and 90%, respectively. The results showed the robustness and the effectiveness of the proposed models.

Keywords: Arabic speech recognition, sentiment analysis, deep learning, COP28.

الخلاصة

في السنوات الأخيرة، لعبت دولة الإمارات العربية المتحدة دوراً محورياً في تعزيز أجندة المناخ العالمي من خلال استضافة أحداث مهمة مثل مؤتمر الأطراف COP28. يعد COP28 بمثابة خطوة هامة للحوار الدولي والتعاون بين الدول لمعالجة تغير المناخ وتسريع الجهود المبذولة للتخفيف من آثاره. في عصر يتسم بالتقدم التكنولوجي السريع، يبرز تطوير أنظمة التعرف على الكلام باللغة العربية كخطوة أساسية في تعزيز إمكانية الوصول والكفاءة وسهولة الاستخدام عبر مختلف المجالات. على الرغم من التقدم الكبير في التعرف على الكلام للغات مثل اللغة الإنجليزية، لا تزال هناك تحديات في تكيف هذه التقنيات بشكل فعال لاستيعاب الخصائص الفريدة للغة العربية. وفي هذا السياق، يعد استكشاف التعرف على الكلام العربي في إطار مؤتمر المناخ الثامن والعشرين بمثابة دراسة حالة خاصة. يستخدم هذا البحث تقنيات التعرف على الكلام في COP28 حيث يعمل على تبسيط الاتصالات، وتسهيل التعاون متعدد اللغات، وتعزيز إمكانية الوصول للمندوبين الناطقين باللغة العربية وأصحاب المصلحة. من خلال التحقيق الشامل في نماذج التعرف على الكلام المختلفة، بما في ذلك CNN، وBI-LSTM، والبنى الهجينة مثل CNN-BI-LSTM وGRU-BI-LSTM، يمكن الحصول على رؤى قيمة حول أدائها وفعاليتها ضمن نطاق العمل. يوفر تحليل المقاييس الرئيسية مثل الدقة والاستدعاء ودرجة F1 عبر فئات المشاعر المختلفة - الإيجابية والسلبية والمحايدة - فهماً دقيقاً لنقاط القوة والقيود في كل نموذج. يأتي النموذج الهجين الذي يدمج GRU وBI-LSTM في المرتبة الأولى، ويعرض أداءً متميزاً بمعدل دقة يصل إلى 94%. وتأتي في المرتبة التالية تقنية GRU المستقلة، حيث حققت دقة تصل إلى 93%. وبعد ذلك، يحذو كل من طرازي CNN-BI-LSTM وCNN حذوهما بدقة تبلغ 91% و90% على التوالي. لقد أثبتت النتائج مدى فعالية الأنظمة المستخدمة.

الكلمات الأساسية: التعرف على مشاعر الكلام العربي، تحليل مشاعر الكلام، التعلم العميق، مؤتمر المناخ

Dedication

To my parents, I cannot express how much I appreciate them. Their comments reverberate in my thoughts and feelings, and their presence endures despite our geographical separation. My motivation to continue my studies and earn a master's degree comes primarily from them. Through my master's journey, my parents, sisters, and brothers have provided me with constant support, encouragement, and understanding for which I am incredibly grateful. I was able to overcome challenges and achieve my goals because of their love, concern, and encouraging comments.

Acknowledgements

I extend my gratitude to the entire community at The British University in Dubai for granting me the opportunity to delve into the field of informatics, which has enriched my knowledge immensely.

I am also grateful to Dr Manar Alkhatib for overseeing this dissertation and facilitating its completion. Additionally, I extend my thanks to Prof. Khaled Shaalan for his guidance and support throughout this research endeavour. Their assistance has been instrumental in producing a research paper of this calibre.

Finally, I would like to express my heartfelt appreciation to my family, especially my sisters, husband, and parents, for their unwavering encouragement and support in my educational pursuits.

Table of Contents

Chapter 1: Introduction	1
1.1 Problem Definition.....	7
1.2 Research Objectives.....	10
1.3 Dissertation Contribution.....	11
1.4 Dissertation Organisation.....	11
Chapter 2: Related Studies	13
2.1 A Review of Natural Language Processing	13
2.2 Sentiment Analysis	14
2.2.1 Sentiment Analysis on Social Media	14
2.2.2 Sentiment Analysis Techniques.....	15
2.2.3 Sentiment Analyses in the Arabic Language	17
2.3 Speech Sentiment Analysis.....	19
2.3.1 Arabic Speech Sentiment Analysis.....	20
2.4 Summary	22
Chapter 3: Methodology.....	23
3.1 Dataset Collection	23
3.2 Audio Sentiment Analysis Categorisation	25
3.3 Preprocessing of the Dataset.....	30
3.5 Feature Extraction.....	35
3.7 Deep Learning Models.....	37
3.7.1 Convolution Neural Network.....	38
3.7.2 Bidirectional Long Short-Term Memory	43
3.7.3 Gated Recurrent Unit	45

3.7.4	The Hybrid Convolution Neural Network-Bidirectional Long Short-Term Memory Approach.....	47
3.7.5	Hybrid Gated Recurrent Units -Bidirectional -Long Short-Term Memory.....	49
3.8	Evaluation Report	50
3.9	Summary	53
Chapter 4: Implementation and Experiment.....		54
4.1	Deployment Environment.....	54
4.2	Experimental Architecture	55
4.2.1	Gathering and Preparing the Dataset	55
4.2.2	Models Utilised in Arabic Speech Recognition.....	56
4.3	Results and Discussion	69
4.4	Conclusion and Future Work.....	89
4.5	Summary	92
References		93

List of Tables

Table 1: The Dataset Characteristics	25
Table 2: Dataset Sample of the Collected Speech	26
Table 3: The Dimensions of Emotional Data Within the Audio Files in the Dataset.....	27
Table 4: The Number of Audio Files	27
Table 5: The Labelling and Numerical Representation of the Sentiments	34
Table 6: The File Name Labelling	35
Table 7: Confusion Matrix Polarity Categorisation.....	53
Table 8: The Hyperparameter Setup for the CNN Model.....	56
Table 9: The Hyperparameters' Setup of the BI-LSTM	59
Table 10: GRU Hyperparameters	62
Table 11: The CNN-BI-LSTM Hyperparameters.....	64
Table 12: The Hyperparameters of the GRU-BI-LSTM Hybrid Model.....	66
Table 13: The Performance of BI-LSTM	69
Table 14: The CNN Performance	73
Table 15: The GRU Performance	76
Table 16: The CNN-BI-LSTM Performance	79
Table 17: The GRU-BI-LSTM Hybrid Model Performance	82
Table 18: Accuracy in the Classifiers	86

List of Figures

Figure 1: Methodology Framework	24
Figure 2: The Distribution of Sentiments on the Datasets	28
Figure 3: The Data Preprocessing for Arabic ASR	33
Figure 4: The Neutral Waveform.....	36
Figure 6: The Positive Sentiment Signal	36
Figure 5: The Signal of Negative Sentiment.....	36
Figure 7: Splitting the Dataset	38
Figure 8: The CNN Structure.....	40
Figure 9: The CNN Architecture	41
Figure 10: The Spectrogram Signal	42
Figure 11: The BI-LSTM Architecture.....	44
Figure 12: The Structure of a Single Cell of the GRU	47
Figure 13: The CNN-BI-LSTM Architecture	48
Figure 14: The GRU-BI-LSTM Structure	50
Figure 15: The performance of the BI-LSTM	70
Figure 16: The Confusion Matrix for the BI-LSTM.....	71
Figure 17: The CNN Performance	73
Figure 18: The Confusion Matrix for the CNN Model.....	75
Figure 19: The GRU Performance	76
Figure 20: The GRU Confusion Measures	78
Figure 21: The CNN-BI-LSTM Performance.....	80
Figure 22: The CNN-BI-LSTM Confusion Matrix	81
Figure 23: The GRU-BI-LSTM Performance.....	83
Figure 24: The Confusion Matrix of the GRU-BI-LSTM Hybrid Model	84

Figure 25: The Accuracy of the Classifiers in Arabic ASR.....	86
---	----

Chapter 1: Introduction

Smart cities are primarily conceived with the objectives of enhancing the quality of life for residents and fostering economic development, with an emphasis on sustainable practices and the effective delivery of services (Angelidou, 2015). Within the framework of smart city development, intelligence gathering entails the continuous monitoring and evaluation of events, inhabitants, and various assets within the urban environment. Subsequently, the collected intelligence is leveraged to improve the efficiency of urban operations, including forecasting the demand and nature of resources, optimising municipal services, and facilitating citizen-centric revisions to policies and regulations (Xiao et al., 2022). A specific challenge in this domain pertains to the development of a scalable model capable of efficiently acquiring contextual real-time data related to residents, assets, or events while operating within resource constraints. This study introduces a novel and sustainable technological approach for the real-time monitoring of the evolution of public opinion within smart cities.

Technological advancements have revolutionised our ability to communicate, access information, and accomplish tasks with greater speed and efficiency. With the introduction of Web 3.0, individuals now have the platform to articulate their thoughts and perspectives on various aspects of their daily lives (Cheng, 2024). This surge in individual expression has garnered interest from businesses, governments, and individuals seeking insights into public sentiment regarding goods or services (Saber & Saad, 2017). However, deciphering this vast and often disorganised data, which also contain private information, presents several challenges. Consequently, fields like sentiment analysis and big data analytics have increasingly focused on addressing these challenges (AlKhatib et al., 2020).

Public opinion encapsulates the collective views, desires, and preferences of the majority of a population on various issues, be they political, commercial, social, or otherwise

(Dong and Lian, 2021). Timely insights into people's perspectives and choices are crucial for businesses aiming to uphold or improve the status of their brands or products, as well as their overall reputation. This information also helps businesses identify and forecast profitable trends in areas like leisure, wellness, and lifestyle. Incorporating such data intelligently into product or service development enables businesses to stay competitive or gain a competitive edge. Governments also frequently base policy decisions on prevailing trends and reactions in public opinion, and monitoring such opinion is instrumental in identifying extremism and predicting and responding to potentially problematic trends and events. In extreme cases, negative public opinion on a particular issue can escalate to civil disobedience or even riots, while unchecked expressions of extremism have been linked to violent incidents, such as those carried out by individuals who have published radical statements online before committing terrorist acts (Ware, 2020). By tracking and assessing public sentiment and opinion in real-time, governments can gather valuable insights to inform decision-making and support the anticipation, mitigation, and prevention of undesirable events, indicating areas of mounting pressure or where intervention may be necessary. Monitoring public opinion is thus recognised as a valuable tool in emergency preparedness, violence prevention, and the maintenance of public order in general.

Sentiment analysis is the study of people's thoughts or feelings about a situation, a discussion on a particular subject, or something else entirely. It has uses in many different fields, such as comprehending human viewpoints via interactions with one another. It is crucial to identify the speakers and understand their speech for a machine to be able to infer human emotions from talks (Dhouib et al., 2022). Thus, researchers first implement speech and speaker identification systems and then do sentiment analysis that has been retrieved. It was possible to determine the sentiment polarities of the sentences from the speech and categorise them as positive, neutral, or negative. Comprehending human emotions can prove

to be immensely advantageous in many situations. As an illustration, computers that are programmed to identify and react to non-verbal human indicators like emotions could customise their responses to match each user has tastes and requirements (Shaalan et al., 2018).

Recent years have witnessed a surge in interest in sentiment analysis as a useful method for deriving conclusions from massive amounts of raw data. Due to the urgent need for these skills, academics have created techniques to measure public opinion in response to this information need (Medhat et al., 2014). By obtaining a deeper grasp of current sentiments on a certain good or service, businesses and organisations can use the insightful information this analytical technique provides to realign their future goals. Enhancing the customer experience is a result of hearing what customers have to say. Businesses show that they care about the needs of their clients when they aggressively seek out and address their complaints. As a result of the loyalty and trust this fosters, retention rates are higher and positive word-of-mouth recommendations are generated. Sentiment analysis is not only useful in business contexts but also has potential uses in the medical field, especially in the management of psychological issues, particularly when dealing with instances that require the diagnosis of depression (Latif et al., 2021).

The increasing number of investigations on sentiment analysis and machine learning proves its significance. From the beginning of the millennium, sentiment analysis within the domain of natural language processing (NLP) has developed as one of the supremely productive fields of study (Liu, 2022). Nonetheless, studies, instruments, and resources devoted to languages other than English – especially Arabic – remain sparse. Furthermore, English language datasets have received most of the attention. Building strong classifiers is a critical step for many NLP and machine learning applications, and training on large datasets

is seen to be a prerequisite for this process (Alkhatib and Shaalan, 2017). Having access to a variety of small, large, and massive databases is quite beneficial for academic study.

Arabic is the sixth most spoken language in the world in terms of speakers, about 380 million native speakers ("Arab world - total population 2011-2021"). Moreover, it is ranked fourth among the languages people use the most on websites (Ray and Shaalan, 2016). Compared to other languages, Arabic has seen comparatively little research done in South Africa, despite its considerable usage. This is mostly because Arabic has intricate grammar and a rich morphology, which makes analysis difficult (Alkhatib and Shaalan, 2017). An Arabic speech is defined for the purpose of this study as an online opinion, post, tweet, blog entry, or commentary that represents the sentiment about climate change. Therefore, our goal is to improve sentiment analysis such that it can reliably identify Arabic sentiment (whether positive, negative, or neutral). The increasing presence of Arabic speech in social networking sites and other digital communication channels has brought attention to the demand for advanced techniques to accurately interpret emotion in Arabic audio. Arabic dialects and Modern Standard Arabic are two examples of the different types that make up the Arabic language (Shaalan et al., 2018) .

Automatic speech recognition (ASR) is a vital aspect of digital signal processing, encompassing elements from linguistics, machine learning, phonetics, and statistics. It is an example of technology that uses software to understand vocal and human speech terminologies (Al-Dujaili and Ebrahimi-Moghadam, 2023). Over recent years, an increasing number of linguists and researchers in speech-related fields have shifted their attention towards ASR techniques. This interest is seen by their integration into several fields, including robotics, transcription, healthcare, and education (Flemotomos et al., 2022). ASR is presently used in various applications due to its effectiveness and efficiency as a result of the technology has swift advancement; for instance, as applied in transcribing software, offering

significant accessibility benefits to Personal Computer users. Voice-activated instructions and portable device searches are two additional noteworthy applications (Cummins and Schuller, 2019). Furthermore, voice recognition technologies are applied to translate speech between different languages. However, ASR is regarded as a challenging issue in the realm of signal processing, requiring multiple processing layers to reach optimal accuracy.

Existing techniques have several drawbacks. First, they offer a compiled and survey-based representation of emotions. For instance, a study (Karami et al., 2018) analysing economic concerns during the 2012 US elections was conducted in 2018, highlighting the delayed nature of the analysis. Furthermore, these methods do not allow for the real-time tracking of opinion evolution and sentiment development, which is essential for modifying strategy in response to public reactions to ongoing events. Finally, there are issues with efficiency and scalability due to the fact that it takes millions of social media postings to get valuable data, which makes them useless for real-time tracking.

COP28 is the abbreviation for the 28th Conference of the Parties and holds immense significance for the United Arab Emirates (UAE) and the universal community at large. As a crucial forum under the United Nations Framework Convention on Climate Change (UNFCCC), COP28 assembles countries from all over the globe to present pressing ecological concerns and chart a collective course towards an eco-friendly and carbon-free society ("COP28 UAE - United Nations Climate Change Conference," 2023.). The UAE hosting COP28 exhibited a unique chance to showcase its obligations to ecological stewardship, and renewable progress on the global stage. As a nation located in a region vulnerable to the effect of atmospheric warming, the UAE recognizes the importance of proactive action to mitigate greenhouse gas emissions, protect natural resources, and build endurance to climate-related risks. Hosting COP28 provides the UAE with a platform to demonstrate leadership in weather protection and innovation, highlighting its efforts to

transition towards a non-polluting economy and promote alternative power resources. By convening COP28, the UAE underscores its dedication to fostering worldwide cooperation and driving ambitious climate commitments to introduce the urgent threat of weather alteration (“We need transformative policies and actions to stand a chance of winning the race to net zero,” 2023). Furthermore, COP28 holds significance for the world as a whole, serving as a critical juncture for universal weather discussions and policymaking. At COP events, countries come together to negotiate and finalise agreements aimed at decreasing emissions, adapting to the atmosphere effect, and mobilising finance to support weather action, such as the Paris Approval. COP28 offers a chance for nations to reaffirm their commitments to the Paris Settlement and accelerate efforts to restrict global warming to well below 2 degrees Celsius above pre-industrial levels, as outlined in the agreement goals. Additionally, COP28 provides a platform for dialogue, knowledge sharing, and partnership via governments, civilian populations, businesses, and other shareholders to address the sophisticated issues posed by weather change (“COP28 - Carbon Market Watch,” 2023.). Ultimately COP28 represents a crucial milestone in the universal fight against climate change, with the potential to shape the trajectory of the worldwide climate agenda for years to come. By engaging in constructive dialogue, forging partnerships, and committing to ambitious climate targets, countries can work side by side to safeguard the planet and secure a sustainable future for generations to come (Coppin and Sander, 2021).

This dissertation explores the complicated concepts and methodologies involved in conducting the sentiment analysis of Arabic speech posts, aiming to examine sentiment patterns in social media data. The research employed OP28, also known as the UN Climate Change Conference, held from November 30, 2023, to December 12, 2023, in Dubai, United Arab Emirates, as a case study for climate change (“COP28 UAE - United Nations Climate Change Conference,” 2023). The objective of this research is to develop a framework that

enables the resource-efficient and sustainable real-time analysis of high-impact networking content, addressing the limitations mentioned earlier. The study introduces an Arabic dataset encompassing both dialect Arabic and Modern Standard Arabic, sourced from YouTube videos. Given the complexities and rich morphology of the Arabic language, language analysis poses challenges, necessitating enhancements in speech sentiment analysis for discerning Arabic speech sentiment tones (positive, negative, or neutral). While the primary focus lies in sentiment analysis applications, the dataset holds potential for various NLP and deep learning tasks. The effectiveness of Arabic speech sentiment analysis is evaluated using five classifiers: Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BI-LSTM), Gated Recurrent Unit (GRU), hybrid CNN-BI-LSTM, and hybrid GRU-BI-LSTM models.

1.1 Problem Definition

Sentiment analysis is essential for understanding the opinions and remarks of people. Its importance is especially noticeable when it comes to the Arabic voice on social media, where it assists in recognising possible issues, gauging customer satisfaction, and guiding significant market choices. Furthermore, the characteristics of the Arabic language, syntactic background, and the lack of extensive, labelled datasets present difficulties for sentiment analysis in Arabic web-based interactions. Developing a robust sentiment analysis model tailored for Arabic audio comments on social networks is considered a challenge. To extract precious insights from customers' experiences and expertise, strategies must accurately categorise these sentiments into positive, negative, or neutral comments (Dhouib et al., 2022). The main goal is to make web-based interaction easier as businesses work to track and understand customer feedback, spot patterns, and take proactive measures to improve their policies.

The expansion of Arabic audio content across various social media and communication platforms underscores the urgent need for advanced techniques to accurately analyse the sentiment conveyed in Arabic speech (Dhouib et al., 2022). Consequently, there is a critical demand to increase our understanding of human influence and enhance the robustness and efficacy of methods employed to address the challenges inherent in speech sentiment analysis.

The field of smart city analytics comprises the monitoring, analysis, and assessment of opinions and emotions expressed through social media platforms. In recent years, the sentiment analysis of social networking postings has gained popularity in learning how the public views current events. Various forms of sentiment analysis have been employed in the current study of this field. Although these methods allow academics to gain a historical understanding of public sentiment, they do not allow for a real-time assessment. Moreover, they are not easily scalable and require a huge number of posts – millions – to be analysed to allow for a more thorough assessment. This research aims to address these limitations by providing a framework that enables a real-time assessment of the changes in the viewpoints held by public figures, or high-impact posts.

Within the field of sentiment analysis, the focus has mostly been on textual data in languages that are commonly spoken, ignoring the sensitivity and complexity of emotion expressed in audio, especially in languages such as Arabic. COP28 is a significant worldwide gathering that addresses climate change, and it provides a useful case study to recognise the dynamics of sentiment in Arabic audios. Nevertheless, the overall insight extraction from Arabic voice sentiment analysis is restricted by the lack of effective machine-learning models designed for this purpose.

The essential task is to build a deep learning model that can effectively identify sentiments from Arabic speeches presented throughout COP28. These demands overcoming several challenges:

- 1) *The sophisticated Arabic morphology*: Because Arabic has such a large vocabulary, sophisticated grammar, and a variety of dialects, it is more challenging than English or other languages.
- 2) *Arabic voice datasets*: It is quite difficult to obtain a large and varied set of Arabic speech datasets derived from opinions expressed on social media about COP28. The lack of sufficient datasets limits the creation of efficient sentiment analysis systems, as well as the model's training and assessment. To guarantee the robustness of the model-building process, a considerable dataset covering comprehensive human emotions – ranging from positive, negative, to neutral – must be gathered and well labelled.
- 3) *Polarity sentiment identification*: Determining the polarity of human feelings requires an expert skilled in sentiment analysis, requiring time and money.
- 4) *The cultural diversity of the Arabic language*: Social customs and principles have a big influence on sentiment analysis for the Arabic language. The traditions and customs of various societies influence people's methods of interaction. Accurately assessing emotions and preventing misconceptions need an awareness of and consideration for these cultural variations.
- 5) *Arabic polysemy and homonymy*: Arabic phrases display homophones and polysemy, suggesting that a word's meaning can change according to its context. It is difficult to distinguish between neutral, negative, and positive feelings because of this ambiguity.
- 6) *Domain adaptation*: Addressing the domain-specific issues raised by COP28 statements, such as the usage of context-dependent sentiment expressions, cultural

allusions, and technical terms. A critical component is maintaining generality while adapting models to this domain.

It is critical to overcome such challenges to fully utilise the potential of Arabic speech sentiment analysis and provide sophisticated insights into the attitudes, opinions, and feelings of the public, as stated at significant global events such as COP28. The successful improvement of a robust deep learning framework designed for Arabic sentiment analysis could equip policymakers, academics, and stakeholders with valuable information for informed decision-making and effective communication strategies in presenting pressing global issues.

1.2 Research Objectives

This dissertation's main objective is to present an extensive framework created especially for assessing the sentiment analysis of Arabic speech, with a focus on the emotions expressed during COP28. The methodology aims to give reliable sentiment classifications for Arabic social media and tackle the difficulties related to sentiment analysis in Arabic speech. The following are the research's main objectives:

- 1) *Generating an extensive labelled dataset related to COP28*: Gather a large and diverse annotated dataset with a range of audios and user satisfaction levels for Arabic speech sentiment analysis, especially for COP28. The sentiment analysis model's performance will be assessed and trained using this dataset as the main source of data.
- 2) *Establishing Arabic speech-specific deep learning sentiment analysis models*: Use sentiment analysis techniques based on deep learning that is in line with Arabic's linguistic characteristics.
- 3) *Analysing and evaluating sentiment analysis algorithms for Arabic speech*: Utilise evaluation criteria such as the F1 score, recall, accuracy, and precision. Furthermore,

use the annotated dataset to create confusion matrices for the suggested framework.

Compare the results to determine the effectiveness of the suggested approaches.

Through the achievement of the previously described research goals, this work aims to enhance Arabic speech sentiment analysis, specifically in the context of COP28 in social media discussions. Obtaining an extensive awareness of Arabic emotional expressions is important for human–machine interactions in smart cities. Thus, there is an urgent need for research projects aimed at creating reliable, culturally aware, and linguistically varied Arabic speech sentiment analysis systems. By bridging the current gap, this project will improve the field of sentiment-aware Artificial Intelligence (AI).

The following research questions are the focus of the dissertation investigation:

RQ1: How can the Arabic speech sentiment analysis deep learning algorithms be effectively categorised?

RQ2: Can sentiment analysis in Arabic speech evaluations be improved by machine learning or deep learning models?

RQ3: What are the percentages of the sentiments (positive, negative, or neutral)?

RQ4: Which methods are most beneficial for recognising Arabic speech sentiment analysis?

RQ5: What is the best accuracy percentage for the sentiment classification of Arabic speech?

1.3 Dissertation Contribution

The dissertation's main goal is to close the knowledge gap in Arabic speech sentiment analysis on social media platforms, with a focus on Standard Modern Arabic and dialect Arabic in the context of the COP28 representation of climate change. To attain superior success in Arabic-language social media conversations, sentiment analysis deep learning classifiers will be implemented, explained, illustrated, and evaluated.

1.4 Dissertation Organisation

The dissertation is structured as follows:

- *Chapter 1: Introduction:* Provides a review of the use of deep learning algorithms for the sentiment evaluation of Arabic text and speech on Internet social platforms, along with information on the problem definition, research questions, objectives, and contribution of the dissertation.
- *Chapter 2: Related Studies:* A full assessment of the research on sentiment analysis, covering topics like sentiment polarities and social media platforms, may be found under related studies. Furthermore, included in this chapter is research on Arabic voice sentiment analysis alongside Arabic text. Lastly, the conclusion provides a quick summary of the main points of this chapter.
- *Chapter 3: Methodology:* Gives a comprehensive explanation of a dataset corpus, covering the procedures for data collection, this involves applying deep learning methods and combining multiple deep learning strategies on the collected data. A brief overview of the major topics covered concludes the chapter, including the evaluation metrics.
- *Chapter 4: Implementation and Experiment:* A deconstructed explanation of the research setup in this section of the dissertation, including the procedures and library resources that were used, a summary of the investigation's settings is suggested in this chapter, along with the initial processing, deep learning, and hybrid deep learning models' hyperparameter configurations. A conclusion for the dissertation is proposed in this section. In addition, this chapter closes with a summary of the directions for future study.

Chapter 2: Related Studies

This chapter covers several subjects, including sentiment analysis on social media and how it is used, as well as a knowledge of NLP with a particular emphasis on works on Arabic speech. It then looks at other approaches to sentiment analysis, including machine learning and deep learning methods. There will then be a study of feature extraction methods and data preprocessing methods.

2.1 A Review of Natural Language Processing

Regarding big data, the majority, about 95%, exists as unstructured texts sourced from websites, emails, and similar sources. To fully utilise the benefits of this massive data size, novel scientific tools, and methodologies are necessary (Cai, 2021). NLP techniques have facilitated this task by efficiently processing large amounts of textual data. NLP can be seen as a subset of AI, employing computational algorithms like machine learning and deep learning to comprehend, learn from, and generate human-like language with semantic coherence (Cai, 2021). NLP has demonstrated its ability to achieve comparable outcomes to those attained through traditional qualitative text analysis methods, as evidenced by studies such as the one conducted by (Guetterman et al., 2018). This study indicated that while NLP successfully identified major themes through traditional text analysis, it struggled to capture nuances and fell short in providing contextual details. NLP expanded in multiple fields and has found applications in healthcare, clinical settings, social media, and various other domains. Despite the significant integration of deep learning techniques into NLP, their adoption in clinical practice remains limited due to their lower interpretability compared to traditional machine learning methods (Casey et al., 2021). Lexicon-based approaches, a subset of NLP techniques, gained popularity in sentiment analysis, particularly during the years 2016–2017. However, recent advancements in sentiment analysis have witnessed a shift from pure NLP methods to conventional machine learning classifiers and deep learning-based

models, which have demonstrated superior performance and achieved state-of-the-art results (Prastyo et al., 2020). Some may argue that sentiment analysis algorithms frequently have trouble understanding complex or unclear speech, especially in situations where irony, humour, or cultural allusions are common.

Over the past decade, Arabic and its dialects have garnered attention in the realm of NLP research. A multitude of studies have delved into different aspects of processing this language and its dialects, encompassing machine translation, resource development, and morphological analysis. Several studies have been proposed to categorise works dealing with Arabic and to elucidate its linguistic characteristics (Alkhatib and Shaalan, 2017; Shoufan and Alameri, 2015).

2.2 Sentiment Analysis

2.2.1 Sentiment Analysis on Social Media

The availability of big data necessitates automated analysis to derive insights. Sentiment analysis encompasses methodologies that automatically categorise human opinions conveyed in a text, including the sentiments and emotions expressed by individuals on diverse topics such as product reviews or views on energy sources (Keshavarz and Abadeh, 2017). It may involve tasks like categorising sentiments into Positive, Negative, or Neutral classes or delving into finer distinctions like extremely positive, positive, neutral, negative, and extremely negative categories (Ahmad et al., 2017), as well as identifying emotions like "Anger", "Sadness", "Fright", and "Happiness". Sentiment analysis operates at three levels: document, sentence, and aspect. NLP techniques are frequently utilised for automated text analysis. Document-level analysis aims to discern the sentiment of an entire document, while sentence-level analysis provides a more granular examination of individual sentences. The aspect-level analysis identifies the sentiment associated with specific attributes often discussed in product reviews (Prastyo et al., 2020). The use of sentiment analysis may raise

ethical questions, especially in situations where privacy may be invaded or where discrimination based on inferred feelings may occur.

Social media platforms, which have accumulated millions of users globally, offer individuals the opportunity to express their views on various topics of interest. As of 2021, tweets are limited to 280 characters, necessitating users to briefly summarise their opinions within this constraint. Analysing Twitter text presents challenges due to data noise, such as spelling errors, grammatical inconsistencies, and the use of emoticons or slang that develop over time (Ahmad et al., 2017). Sentiments deduced through Twitter sentiment analysis have numerous applications, including identifying correlations between the expressed sentiments and polling data during the 2016 presidential election (Joyce and Deng, 2017), evaluating reviews of the tourism industry (Ramanathan and Meyyappan, 2019) and the pilgrimage season (Ottom Nahar, 2021), and gauging public emotions during the COVID-19 pandemic (Weger et al., 2023). Sentiment analysis, also referred to as opinion mining, amalgamates various NLP techniques such as tokenisation, stemming or lemmatisation, and bag-of-words with machine learning models to compute sentiment scores for terms within a given text (Kanakaraj and Guddeti, 2015). These sentiment scores, in turn, enable the determination of the overall sentiment polarity of the text, which may be positive, neutral, or negative. Critics could draw attention to the shortcomings of sentiment analysis methods that only include textual data, ignoring non-verbal clues like body language, tone of speaking, and facial expressions that can also be significant sources of sentiment signals.

2.2.2 Sentiment Analysis Techniques

Machine learning methodologies comprise three main categories. Firstly, unsupervised learning involves analysing words within documents, where common words are grouped into multiple clusters predefined by researchers. Secondly, semi-supervised learning employs both unlabelled and a limited number of labelled data during classifier training.

Here, unlabelled data extract crucial features, while labelled data refine the classifier output (Ligthart et al., 2021). This study focuses on investigating the third category: supervised learning.

Numerous studies are currently underway focusing on sentiment analysis utilising diverse methodologies. Among the pioneering efforts in employing machine learning for sentiment analysis on internet film reviews is the research conducted by (Pang and Lee, 2008). Their comparative analysis with human assessment revealed that support vector machines and Naive Bayes, in particular, could effectively discern emotions from movie reviews. Significant studies have endeavoured to enhance sentiment analysis techniques, with one approach involving supervised machine learning classifiers. This method utilises labelled data as features during training phases, enabling learning algorithms to generate the predicted output (Bhavitha et al., 2017). Subsequently, the trained classifier can predict the sentiment polarity of unlabelled data. However, a significant drawback of this approach is its reliance on a substantial amount of labelled data for training, which can be inefficient if manual labelling by humans is required.

Nevertheless, numerous studies have demonstrated that supervised learning outperforms both semi-supervised and unsupervised learning in various aspects (Ligthart et al., 2021). For instance, in one study, a sample of 3,102 tweets related to COVID-19 was tested using machine learning with a Term Frequency - Inverse Document Frequency feature extraction technique, achieving a maximum accuracy of 80% for Support Vector Machine (SVM), 78% for Naive Bayes, and 76% for Logistic Regression (LR) (Amin et al., 2021). Additionally, using machine learning classifiers on a Kaggle dataset of airline reviews resulted in accuracies of 83.31% for SVM, 81.81% for LR, and 73.00% for Naive Bayes (Dutta Das et al., 2017). Certain sentiment analysis methods might be biased, especially when

used in environments with different linguistic or cultural norms. As an outcome, sentiment expressions may be interpreted incorrectly or inaccurately.

Because of the necessity for large, labelled datasets to train machine learning classifiers, numerous studies have turned to hybrid methods in sentiment analysis. These methods combine the polarity obtained from lexicon-based approaches with machine learning training. (Maurya and Jha, 2024) observed that lexicon methods exhibited high precision but low recall, prompting them to hybridise Manufacturer Product Quality Assessment with SVM for enhanced performance. (Rajeswari et al., 2020) determined that integrating lexicon into hybrid approaches addresses the Neutral class classification issue, resulting in overall accuracy improvements when evaluated across product, Twitter, and movie datasets using the Naive Bayes, Linear regression, SVM, and Decision Tree classifiers. Additionally, the authors emphasised the significance of feature extraction and found TF-IDF to yield the highest accuracy. The accuracy and calibre of the datasets used to train sentiment analysis algorithms may come under fire. Results that are skewed and have poor generalisation can be caused by biased or limited datasets.

2.2.3 Sentiment Analyses in the Arabic Language

Arabic is a language that has a strong connection with Muslim religious traditions and is currently an official tongue for millions of people living in Arabic-speaking countries. Arabic is a Semitic language with 28 letters in total. As of 2022, there are approximately 457 million native Arabic speakers worldwide, spread over 22 Arab-speaking countries, or about 5% of the globe's population (“Demographics of the Arab world,” 2024). The population is projected to rise to about 533 million by 2030, according to (“Arab Day for Population and Development October 28, 2023 | HPC,” 2023). Additionally, Arabic is now recognised as the fourth most frequently used language on websites, having experienced a notable surge in popularity (Ray and Shaalan, 2016). However, the quantity of studies on natural language

processing pertaining to Arabic might not match the language's extensive usage (Mozetič et al., 2018). Furthermore, Arabic sentiment analysis methods have not progressed as far as English sentiment analysis approaches have. Nonetheless, research (Alkhatib and Shaalan, 2017) shows that there is an increasing amount of studies dedicated to Arabic sentiment analysis.

To analyse the personality and sentiment of Arabic interactions on social media sites, the researchers (Tartir and Abdul-Nabi, 2017) created a supervised machine learning system. The impact of platform-specific features on performance techniques for managing Arabic dialects, the suitability of standard features from English to Arabic, and the most efficient means of lexical information transmission were the four areas of focus of their contribution. The sentiment analysis and subjective features of four distinct platforms were also investigated: synchronous chat, Wikipedia talk pages, Twitter, and online discussion forums. These platforms make use of a variety of languages, each with its own distinct goal. The lexeme has been demonstrated to be better suited for sentiment analysis, while the expanded tag set, in conjunction with lemma information, is more appropriate for subjectivity analysis (Abdul-Mageed et al., 2014). The representativeness and calibre of the datasets used to train sentiment analysis models may come under investigation. Results that are limited and have poor generalisation can be caused by biased or limited datasets.

Furthermore, the researchers introduced a manually annotated corpus in (Mourad and Darwish, 2013), where each sentence of newswire pieces was annotated individually. This dataset comprises 2,855 sentences and was utilised by the researchers to showcase an automated approach to subjectivity and sentiment analysis. The researchers investigated the impact of different levels of preprocessing frameworks on the task of classifying subjectivity and sentiment analysis, using the F-measure criteria to assess the results. They demonstrated that, in terms of subjectivity classification, combining stemming with morphology and

adjective attributes yielded the highest F-measure of 71.54%. However, for emotion classification, the combination of stemming with morphological and adjective elements achieved the highest score of 95.52%. Some opponents can argue that the assessment metrics used to evaluate sentiment analysis models are inappropriate because they do not adequately account for the complicated nature of emotional expression or performance in the real world.

The gap in research on Arabic sentiment analysis stems from the limited exploration and development of effective techniques tailored specifically for the Arabic language. Despite the growing interest in sentiment analysis across various languages such as English, Chinese, and Russian, there is a noticeable shortage of studies and frameworks dedicated to Arabic. This gap is largely due to the scarcity of adequately labelled datasets designed specifically for Arabic sentiment analysis. While some resources exist, they are often small in scale and may not fully capture the diverse linguistic nuances and social contexts inherent in Arabic speech.

Another aspect contributing to the academic gap is the complication of sentiment analysis in Arabic, which comprises not only Modern Standard Arabic but also varieties of dialects spoken among different areas. Each dialect may exhibit specific linguistic characteristics and expressions of sentiment, requiring special techniques for sentiment analysis.

2.3 Speech Sentiment Analysis

Recent research in speech sentiment analysis has primarily focused on end-to-end methodologies, exemplified by projects such as the Universal Speech Model (Zhang et al., 2023) and Whisper (Radford et al., 2022). These end-to-end deep learning models have significantly enhanced ASR performance by enabling direct learning from the speech waveform, thereby eliminating the need for intermediate feature extraction layers (Wang et al., 2019)(Radford et al., 2022). Whisper, in particular, stands out due to its multitask training

methodology, incorporating voice activity detection, language identification, ASR, and speech translation. It has demonstrated state-of-the-art performance on various benchmark datasets, including TEDLIUM (Rousseau et al., 2012) and LibriSpeech (Panayotov et al., 2015). However, concerns have been raised regarding its susceptibility to adversarial noise (Olivier and Raj, 2023).

2.3.1 Arabic Speech Sentiment Analysis

Systems for recognising emotions in spoken words differ according to the language, the dataset's extracted features, and the applied classifiers. There has been limited research conducted, particularly on Arabic audio, despite numerous studies aimed at identifying emotions from speech across various languages. The KSUEmotions corpus (Meftah et al., 2020), which is acknowledged as one of the most significant Arabic corpora, was used by the researchers (Ali and Hifny, 2020) for Arabic sentiment analysis. The study employed distinct models for emotion identification. One of these models, the hybrid CNN-LSTM-DNN model, demonstrated superior performance compared to other models. It incorporated Convolutional Neural Networks (CNN), Bi-Directional Long Short-Term Memory (BLSTM), and Deep Neural Networks (DNN). Although the tongue of Arabic is a treasure mine of intriguing issues to explore, all the works in the area of Arabic natural language processing have only scraped the surface.

In a separate research endeavour aimed at recognising emotions in Egyptian Arabic speech, sentiment analysis, and emotion identification techniques were employed by (Abdel-Hamid, 2020). The dataset utilised comprised partially natural data. Wavelet, spectral, and prosodic properties were extracted and employed as inputs. The classification methods utilised included both the SVM and K-Nearest Neighbours. Researchers conducted numerous tests, encompassing both speaker-dependent and speaker-independent scenarios. They categorised the results based on each classifier, extracted feature, speaker, gender, and

emotion. It can be difficult to tailor sentiment analysis techniques to particular languages or domains, and detractors might counter that current methods frequently have trouble generalising much beyond their training sets.

In the realm of Arabic ASR, the initial End-to-End model was introduced by (Ahmed et al., 2018), utilising recurrent neural networks combined with Connectionist Temporal Classification. Subsequent advancements have expanded upon this foundation, including the development of transformer-based models that exhibit proficiency in both Modern Standard Arabic and dialect Arabic (Belinkov et al., 2020). One significant challenge facing end-to-end ASR models is the considerable need for annotated data, particularly in languages with limited resources such as various Arabic dialects. To tackle this challenge, self-supervised and semi-supervised learning approaches are gaining prominence. These models, such as XLS-R and WAV2VEC2.0, initially learn valuable representations from extensive amounts of unlabelled or poorly labelled data, which can subsequently be fine-tuned for certain functions (Baevski et al., 2020). A different self-supervised approach, W2V-BERT, applies masked linguistic modelling and contrastive learning techniques. By fine-tuning the FLEURS dataset – which reflects Egyptian dialect-accented standard Arabic – it has been adapted for Arabic ASR. In contrast to whisper, WAV2VEC2.0 and W2V-BERT require a fine-tuning step in order to decode data effectively.

A speech waveform comprises a wealth of information, making it a precious resource. Recognising the optimal collection of data related to sentiment is often a challenge. Consumer applications require techniques that are effective in terms of resource utilisation, which corresponds with the difficulty of identifying the ideal set of sentiment-related information. The limited availability of corpora constrains the broader application of machine learning techniques, particularly when employing deep learning methods, thereby restricting

the diversity of prior endeavours. Recent advancements in sentiment analysis have demonstrated the remarkable effectiveness of these latter approaches.

2.4 Summary

An overview of the basic and recent studies on sentiment analysis is given in this chapter, with a focus on sentiment analysis in the Arabic language. It presents an overview of NLP for the English and Arabic languages. The chapter also covers sentiment analysis on social media, sentiment analysis techniques, and sentiment analyses in the Arabic language. It also emphasises the speech sentiment analysis, particularly for the Arabic language.

Chapter 3: Methodology

Understanding the emotional content of spoken language is necessary to examine the sentiment of dialect speech using deep learning models. This field of AI is significant due to it enabling machines to perceive and react to emotional cues in speech, hence improving machine–human interaction. Arabic speech presents unique challenges due to its unusual phonetic features, diversity of dialects, and cultural nuances, necessitating a specific approach for effectively identifying emotions.

Only a multilayer neural network is used in the suggested way to recognise emotions in Arabic speech. Using the voice input provided to the system, the sentiment emotion recognition system's primary function is to identify emotions in speech. After that, the multilayer neural network carries out the procedure of mining and picking characteristics.

This approach outlines the methodical process of creating a reliable sentiment analysis system for Arabic speakers, specifically focusing on dialect speech. Every stage, from the data collection to model training and deployment, is carefully designed to specifically handle the inherent linguistic diversity and emotional expressions found in the Arabic language. The schematic in Figure 1 illustrates the fundamental elements of the framework.

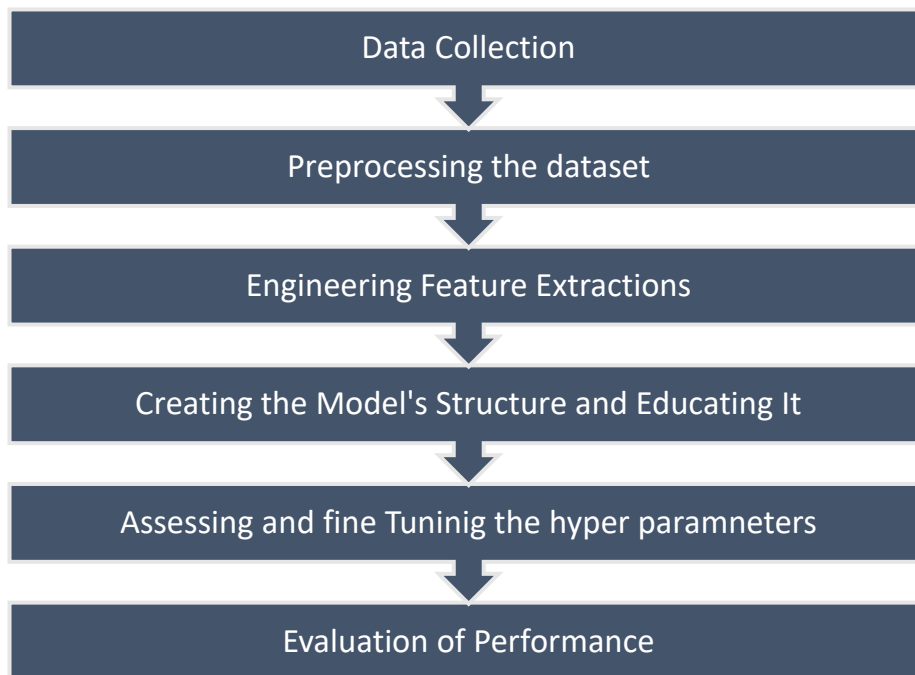
3.1 Dataset Collection

In line with the article (Garouani and Kharroubi, 2022), Arabic dialects and Modern Standard Arabic are mixed on social media sites. These linguistic variations have certain similarities as well as remarkable differences that may impact the application of sentiment analysis. Dialect Arabic, sometimes known as "colloquial Arabic", is used in everyday conversations by Arabic speakers. There are numerous countries and regions where this dialect is commonly utilized. Spoken Arabic accents from the Gulf, the Maghreb, Egypt, and a few more places fall under this group. Online platforms and internet users use dialect

Arabic in writing but with regional variations (Elnagar et al., 2020). Dialect Arabic and Modern Standard Arabic are the main topics of this dissertation.

Figure 1

Methodology Framework



Arabic linguistics has a sophisticated component foundation, multiple morphological rules, and both Templatic and affixational morphemes. Pronouns and related phrases are often used in Arabic, leading to a broad range of inflections. Exceeding the 10,000 origins that form the basis of nominals and tenses in Arabic contributes to the language's complex morphology (Alkhatib and Shaalan, 2017) Arabic morphology comprises 120 patterns that highlight the significance of the 5,000 basic origins. Arabic also allows terms to be placed in a phrase to emphasize different areas, a feature known as variable word order. The grammatical resolver distributes the entry icons provided by the dictionary analyser to determine a sentence structure that complies with Arabic grammatical rules (Shaalan et al., 2018).

3.2 Audio Sentiment Analysis Categorisation

For this research, a variety of expressive speeches for COP28 in the Arabic dialect were gathered from YouTube, Instagram, and X videos. A dataset representing reviews classified into positive, negative, and neutral feelings was created by merging the data. A corpus of 5,744 audios, comprising speech comments and suggestions, was compiled between November 2023 and March 2024. The sample consists of about 60 Arabic utterances in an Arabic dialect. In total, more than sixty Arabic speakers – roughly thirty men and thirty women – between the ages of 25 and 70 contribute more than twelve hours of speech to the samples. The characteristics of this classification are listed in Table 1. Triple categories were created from the evaluations: neutral (NON), negative (NE), and positive (PO). The review speech material was first extracted into an Excel file and then transformed into CSV format for preprocessing because CSV is compatible with several processing programs.

Table 1

The Dataset Characteristics

No. of speech audios	5,744
Start date	Nov 30, 2023
End date	March 25, 2024

The review committee carefully viewed and evaluated each video to pinpoint particular moments that displayed overt emotional indications. The purpose of the research was to evaluate how well different machine learning algorithms detected attitudes. After extracting the audio recordings from the video, they were divided into smaller segments. Each segment was taped in MP3 format, lasting anything from one to three seconds. The speech was divided into segments using the speech's natural pauses, making sure that each segment contained a complete phrase (Meftah et al., 2020). Assigning texts to these specific

categories is the aim of sentiment analysis tools, which also provide valuable insight into the emotional context of the text. These categories make it easier to comprehend comments, opinions, and trends in different texts.

Samples of the Arabic speech transcriptions into Arabic text are presented in Table 2. These emotions therefore functioned as labels for the audio snippets. As part of the data preparation process, steps were taken to eliminate any undesirable disruption, such as noise from the surroundings and music, before categorisation.

Table 2

Dataset Sample of the Collected Speech

Comments	Translate into English	Class	Labels
عرض مؤتمر المناخ حلول نافعة وعملية	The conference presentation in Copenhagen is unforgettable	Positive	PO
على الرغم من الجهود المبذولة في مؤتمر المناخ (COP28)، فإنه من المحبط أن نرى عدم وجود التزام من بعض الدول المشاركة.	Despite the efforts made at COP28, it is disappointing to see the lack of commitment from some participating countries	Negative	NE
حضر المؤتمر العديد من الدول	The conference was attended by many countries	Neutral	NON

The number of files and the size of data collected for the polarities of the audios – labelled Positive (PO), Negative (NE), and Neutral (NON) – are shown in Table 3 and Table 4 respectively. Figure 2 proposes the distribution percentages of the sentiments in the dataset.

Table 3*The Dimensions of Emotional Data Within the Audio Files in the Dataset*

Datasets	Sentiments			Total
	Positive	Negative	Neutral	
Audio files' size	800 MB	32.8 MB	274.4 MB	1.11 GB

Table 4*The Number of Audio Files*

Datasets	Sentiments			Total
	Positive	Negative	Neutral	
No. of audio files	4001	174	1569	5744

****Examining the Emotion Distribution in the Datasets:**

The presented Tables 3 and 4, and Figure 2, offer an extensive breakdown of sentiments across diverse datasets, encompassing positive, negative, and neutral emotion categories, along with the total number of audio files and audio sizes in each category. This analysis sheds light on the sentiment distribution within the datasets, which is pivotal for understanding the composition and features of the data utilised for Arabic speech recognition.

****Positive Emotion: ****

The dataset exhibits a substantial number of audio files associated with positive sentiment, totalling 4,001 audios and 800 MB data size. This abundance of positive sentiment data underscores the prevalence of affirmative, optimistic, or approving expressions within the speech content. The prominence of positive sentiments suggests a significant presence of affirmations, agreements, or supportive remarks in the speech data.

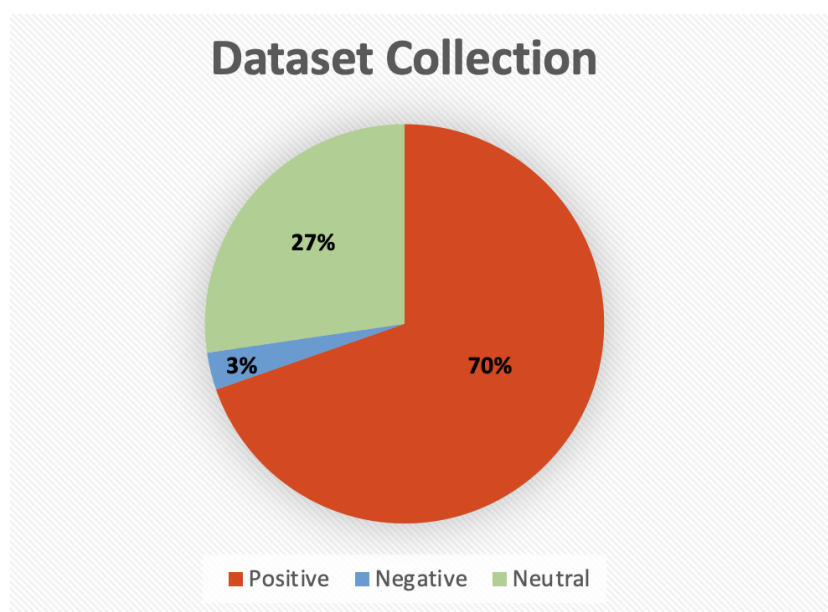
****Negative Emotion: ****

Conversely, the dataset contains a relatively modest number of audio files labelled with negative sentiment, comprising only 174 reviews and 32.8 MB data size. This scarcity

of negative sentiment data highlights a notable imbalance in sentiment distribution, with negative sentiments being less represented compared to positive sentiments. The limited presence of negative sentiments indicates a paucity of content expressing dissatisfaction, criticism, or disapproval within the speech data. It's worth mentioning that the lack of negative sentiment data may be attributed to the nature of the COP28 conference environment, where participants are more inclined towards constructive dialogue, cooperation, and problem-solving rather than expressing negative feedback or criticism openly.

Figure 2

The Distribution of Sentiments on the Datasets



****Negative Emotion: ****

Conversely, the dataset contains a relatively modest number of audio files labelled with negative sentiment, comprising only 174 reviews and 32.8 MB data size. This scarcity of negative sentiment data highlights a notable imbalance in sentiment distribution, with negative sentiments being less represented compared to positive sentiments. The limited

presence of negative sentiments indicates a paucity of content expressing dissatisfaction, criticism, or disapproval within the speech data. It's worth mentioning that the lack of negative sentiment data may be attributed to the nature of the COP28 conference environment, where participants are more inclined towards constructive dialogue, cooperation, and problem-solving rather than expressing negative feedback or criticism openly.

****Neutral Sentiment: ****

The dataset was allocated to neutral sentiment, with 1,569 audio files and 274.4 MB data size. Neutral sentiment signifies a lack of explicit positive or negative affective content, representing an objective or impartial stance in the speech data. The prevalence of neutral sentiments underscores the importance of capturing and analysing non-emotional or used linguistic expressions, which play a crucial role in natural language communication and discourse.

****Total Dataset Composition: ****

In total, the dataset comprises 5,744 audio files and 1.11 GB data size across all sentiment categories, providing a diverse and representative corpus for training and evaluating Arabic speech recognition models. The distribution of sentiments within the dataset reflects the natural variability and complexity of human speech, encompassing a wide spectrum of emotional states, attitudes, and linguistic nuances.

****Implications for Speech Recognition: ****

The analysis of sentiment distribution within the datasets carries significant implications for the development and optimisation of Arabic speech recognition models. The scarcity of negative sentiment data underscores the need for careful consideration and the mitigation of sentiment imbalance during model training and evaluation processes.

Researchers and practitioners must devise strategies to address the imbalance in sentiment

distribution, such as data augmentation techniques, synthetic data generation, or transfer learning from related domains.

Additionally, the observation of sentiment distribution within the COP28 conference context provides insights into the nature of discourse and interaction dynamics prevalent in such environments. The predominance of positive and neutral sentiments reflects a culture of collaboration, consensus-building, and constructive engagement among conference participants. Understanding the unique characteristics of sentiment distribution within specific domains or contexts, such as COP28, enables the development of tailored speech recognition solutions that effectively capture the linguistic nuances and communication patterns inherent to those environments.

In conclusion, the analysis of sentiment distribution within the datasets offers valuable insights into the composition and characteristics of data utilised for Arabic speech recognition. By addressing sentiment imbalance and leveraging domain-specific insights, researchers can advance the development of more robust, contextually aware, and inclusive speech recognition systems tailored to the nuances of Arabic language and communication within diverse domains like COP28 and beyond.

3.3 Preprocessing of the Dataset

Preprocessing is a crucial stage in both sentiment analysis and NLP to ensure higher efficacy in sentiment analysis and to enhance the quality of the generated dataset. Among its primary objectives are text cleansing, normalisation, tokenisation, stop word removal, and stemming. Several steps may be involved in the preprocessing phase, depending on the study's objectives and the chosen language. Research assessing the efficacy of Arabic sentiment categorisation employed unique preprocessing (Abdel-Hamid, 2020). This included emotional support and a customised stop list as strategies. Emojis can be replaced

with custom stop lists and content phrases to improve Arabic emotion categorisation, according to the experiment's findings.

Creating a system for ASR starts with preprocessing, which is the initial stage before further stages of the operation. Building vectors that extract relevant information and differentiate signals produced with vocal cord vibrations from those that are not the major purposes of this. Prior to featuring engineering extraction analysis, the voice signal is preprocessed to make it more acceptable (Zerari et al., 2019).

Development of the feature of the voice data that is received and the optimisation of the process of extracting valuable features are the goals of data preparation in ASR. ASR models need to go through a number of procedures in order to guarantee that the acoustic signals are ready for further examination (Garouani and Kharroubi, 2022). The essential components of Arabic audio data for analysis are as follows:

1. Audio Importing and Transforming: To begin, sound files in the widely utilised .wav scheme for voice information that includes speaker utterances must be imported. The signal is transformed into a 2D NumPy array once the audio files have been parsed, and this allows the sound's amplitude or strength to be seen over time. In this research, the audio file's sample rate will be converted into a sample rate (16 KHz), where the default sample rate is 44.1 kHz, due to its ability to provide more accurate high-frequency data. For audio signals, a sampling rate of 16 kHz is sufficient to capture frequencies up to 8 kHz, which include the relevant frequency range of the human voice. Generally, the 16 kHz sampling rate is a practical selection for ASR, presenting a good balance between capturing relevant acoustic data, arithmetical efficiency, and harmony.

2. Dataset Augmentation: Mechanisms for dataset augmentation are put in place to increase the dataset's circularity and variance. This could mean moving the audio to the left

or right in terms of spatial positioning, or it could mean slightly changing the pitch. With augmentation, the ASR mechanism may learn from a wider variety of entry data.

3. Mel Spectrograms: By altering the sound input, Mel spectrograms are created. To graphically depict the sound waveform, spectrograms divide it up into a range of frequencies across time. Mel spectrograms provide accurate capture of the frequency components of the speech stream, which are crucial data for feature extraction in ASR (Zerari et al., 2019).

4. Mel-Frequency Cepstral Coefficients (MFCCs): A modulation of the voice signal produces MFCCs in addition to spectrograms. MFCCs, which are compact and crucial, are a collection of coefficients taken from the audio spectrum. MFCCs are excellent at properly capturing the fundamental components of human speech.

5. Using Data Augmentation Strategies to Enhance Spectrogram Data: The dataset is further enhanced by utilising spectrograms entirely. Model resilience is sometimes enhanced by the application of techniques such as temporal and frequency masking, as in SpecAugment.

6. Normalisation and standardisation procedures: These are utilised to guarantee that the scales for the entry characteristics are consistent. This is essential for the mathematical framework to converge during training. Using standardisation, which normalises the data bank, the data can be changed to have a mean.

7. Classification: Every audio clip is categorised with the appropriate sentiment to aid in supervised learning.

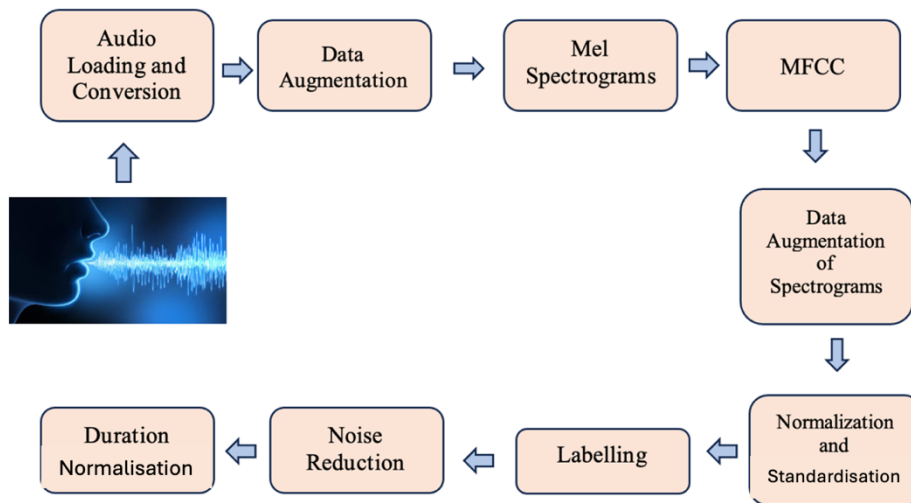
8. Noise Reduction: During preprocessing, the audio file may have background noise, quiet areas, or spurious noises removed. Noise reduction techniques raise the waveform-to-noise percentage, which enhances the technique's voice recognition performance.

9. Duration Normalisation: Since audio files can have varying lengths, it is necessary to normalise the durations to a constant length in order to enable efficient learning from the

data. One frequent strategy to get ready for modelling is to take an average of each audio recording's length (Dhouib et al., 2022). These preprocessing methods improve the quality of the input data ASR, enabling ASR techniques to efficiently extract and identify audio patterns in Arabic tongue utterances.

Figure 3

The Data Preprocessing for Arabic ASR



3.4 Annotation

Annotation has significance due to machine learning and deep learning strategies mainly depending on labelled data for recognising patterns and offering forecasts or categorisations on new, unidentified data. Unreliable findings could emerge from models not being able to perform to their full potential in the absence of exact and extensive annotations in training data (Dhouib et al., 2022).

Since annotation is a prerequisite for training deep learning and machine learning mechanisms, it plays a key role in the field of emotional speech recognition. Using sentiment recognition models, sentiment analysis primarily uses labelled data to identify correlations and patterns in audio waveforms that are suggestive of various emotional states (Zerari et al.,

2019). Based on annotated datasets, these models acquire and apply their knowledge of diverse sentiment cues.

To establish a reliable reference for model training, annotators identify positive, negative, and neutral voice data segments. The model's capacity to precisely recognise and categorise speech emotions is strongly impacted by the degree of precision and thoroughness of this labelling. Fully annotated datasets improve the model's performance by guaranteeing reliable and consistent emotion identification capabilities, even with unidentified voice samples (Abdel-Hamid, 2020).

The first job was to identify and understand the voice cassettes after the dataset was collected. Each dataset had its naming convention, and the emotion label for further categorisation was ascertained from the file names. The labels and their accompanying numerical representations are displayed in Table 5. An example of labelling the filenames is shown in Table 6. The dataset analysis made use of wave charts that displayed sound samples that were chosen at random (Ahmed et al., 2018). The purpose of these plots was to give the properties of the data under inquiry a visual illustration. In order to evaluate the quality of the speech data, the participant's comprehension of the emotions being expressed as well as any possible background noise was measured and detected in the voice recordings.

Table 5

The Labelling and Numerical Representation of the Sentiments

Sentiment type	Labels	Numerical representation
Positive	PO	0
Negative	NE	1
Neutral	NON	2

Annotation in this dissertation refers to the process of assigning certain emotion categories to each audio segment. Based on the file ID name and the acronyms for the various sentiment categories, names were assigned to the audio files. The MFCC_Data_PreProcess

function extracts the first two alphabet characters that denote the emotion type in order to acquire the labels from the filenames. The filenames are utilised as the label after extraction. The emotional content of the audio segments adds the labels to the "labels" list in the data dictionary.

Table 6

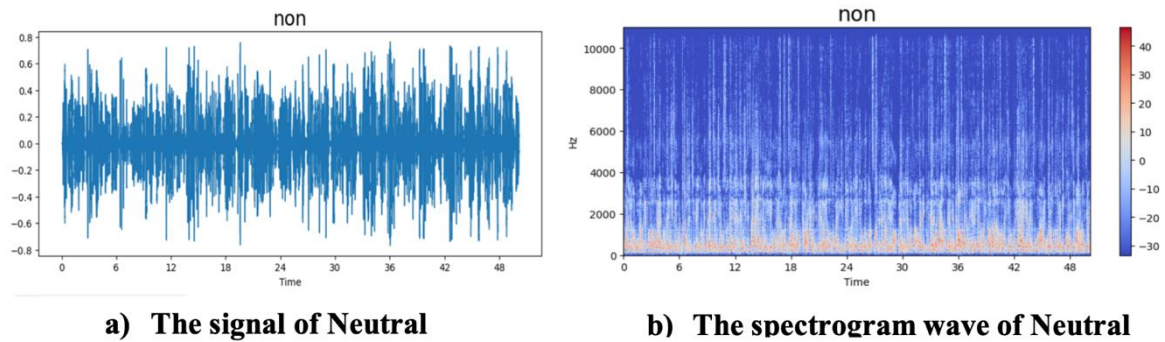
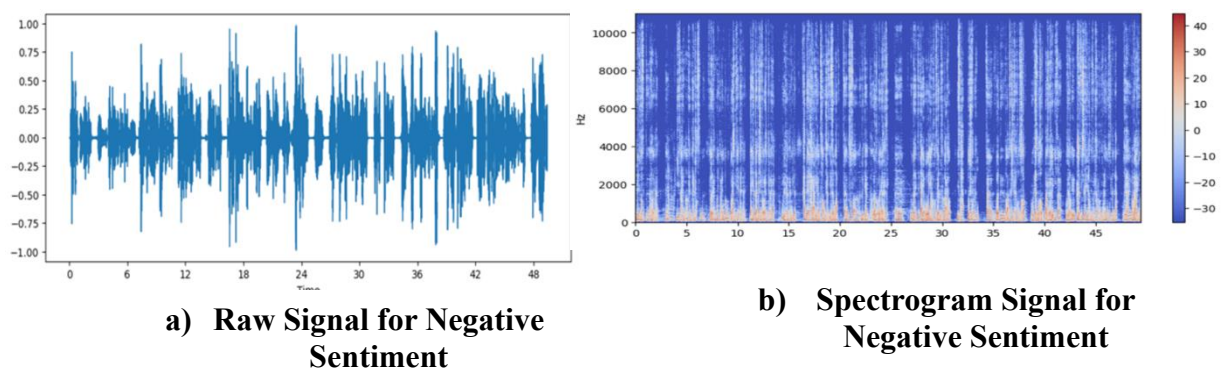
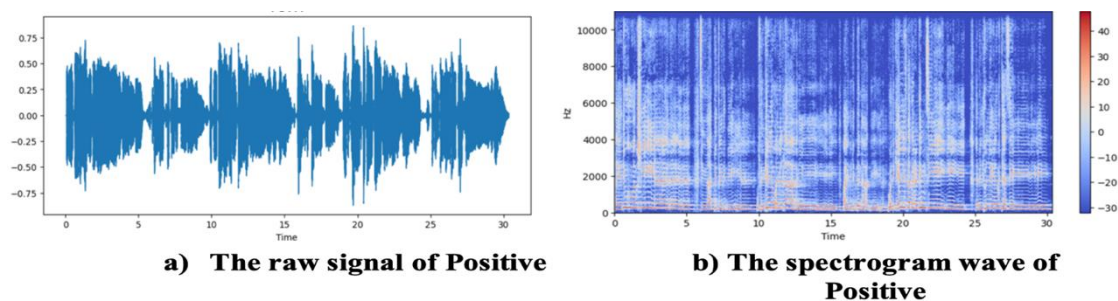
The File Name Labelling

Filenames
PO فرحان_PO.wav
NE نقد_NE.wav
NON غذاء_NON.wav

3.5 Feature Extraction

Multiple components that each communicate a distinct emotional quality make up the speech signal. MFCC is a method that extracts important information from sound segments in order to pinpoint important characteristics for voice analysis. MFCCs are a brief collection of characteristics (usually in the range of 10 to 20) that describe the spectra envelope pattern of an audio source. When calculating the MFCC characteristics, higher frequencies are subjected to linearly diverged frequency filters and lower frequencies to logarithmically spread frequency filters. This design mimics the low-frequency sensitivity found in the human ear, according to the findings of (Saxena et al., 2022).

The following pictures display each audio sample's MFCC matrix calculation. They show the raw, unedited data for each sentiment as well as the spectrogram-visualised MFCC features. Figures 5(a) and 5(b) display, respectively, the signals for the spectrogram wave and the neutral raw data. Figure 6 displays two signals, (a) for the raw data signal for the negative sentiment, and (b) for the spectrogram signal. Figure 7 displays the raw data and the spectrogram wave for positive sentiment in (a) and (b), respectively.

Figure 4*The Neutral Waveform***Figure 6***The Signal of Negative Sentiment***Figure 5***The Positive Sentiment Signal*

3.6 Dataset Splitting

Dividing data collection into training, validation, and testing collections is one series in the machine learning and deep learning operation. It ensures that the system will learn from certain data, adjust its variables in another area, and then be tested in varied fields to assess its effectiveness through this process (Scheurer et al., 2024).

1. Training Set: The model is trained using the dataset's segment, which is made up of the training set. During training, the model finds characteristics and attributes in the data, which helps it learn to forecast. Typically, 70–80% of the dataset is made up of the training set.

2. Validation Set: Adjusting hyperparameters is the responsibility of the validation set, keeping an eye on the mechanism's performance throughout the training. By providing a dataset that was not used during the model's training process, it acts as a safeguard against overfitting. Using data, the validation set helps choose the best-performing model based on effectiveness. Ten to fifteen per cent of the data are often set aside for this use.

3. Testing Set: The testing set is implemented to assess the trained approach's ultimate performance. It provides an impartial examination of the model's effectiveness with fictitious data. Typically, the remaining 10–15% of the data are given to the testing set.

Using a 70:20:10 ratio, the Arabic corpora – either text or speech – that were gathered for analysis were partitioned into training, validating, and testing data collection in this dissertation. A flow chart for the training, validation, and testing data collection for the voice and text reviews is presented in Figure 7.

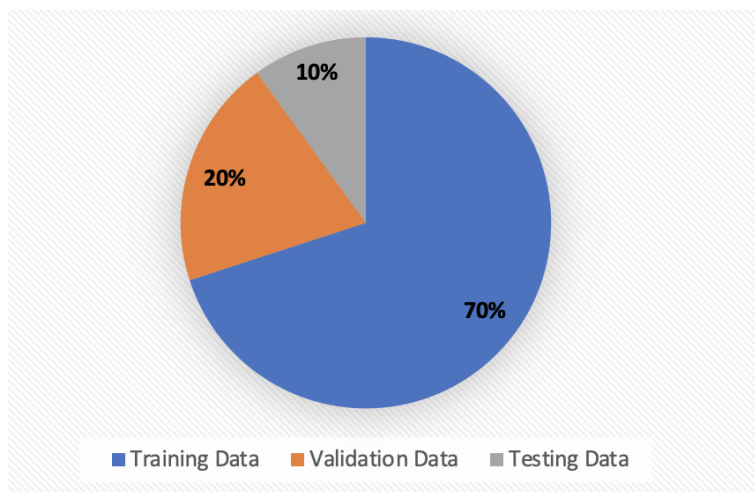
3.7 Deep Learning Models

To find out how satisfied people are with the events happening in the world, it's critical to evaluate people's thoughts about these actions. Currently, COP28 is an event related to climate change all over the world and is a crucial event for the expression of

opinions (“COP28 UAE - United Nations Climate Change Conference,” 2023). As such, data from several social networking platforms were collected and analysed to guide decisions intended to boost people's satisfaction and loyalty. This dissertation offers deep learning approaches designed to analyse a substantial volume of people's opinions on COP28 in Arabic regions by assessing both the positive and negative sentiments stated in these reviews. CNN, BI-LSTM, GRU, and hybrid CNN-BI-LSTM and GRU-BI-LSTM models were used to compare the classification results.

Figure 7

Splitting the Dataset



3.7.1 Convolution Neural Network

With regard to NLP missions, Convolution Neural Network (CNN) is considered one of the more extensively utilised deep learning approaches, and its use yields encouraging outcomes (Gu et al., 2018). To address sentiment analysis challenges, natural language scholars have expressed an interest in using CNN.

A reliable deep learning mechanism for tasks, for instance, computer vision and image identification, is CNNs. The way the brain processes visual data affects CNNs. According to (O’Shea and Nash, 2015), they were specifically designed for recognising

structures and attributes in photos independently. CNNs are computational devices that stimulate the brain's neural networks. Other people can use the three basic levels – the input, output, and hidden layer – while certain systems just employ the layers of input and output. Several scholars examining NLP in conjunction with sentiment analysis have been conducted in the past few years, yielding promising findings.

CNNs are well known for their capacity to extract various feature levels from input data. The layers grasp basic features like edges and patterns, but they also comprehend more sophisticated features like objects and shapes. CNNs function well on issues involving image and video identification because of their structure. It is crucial to note that training a neural network (CNNs) can be resource- and time-intensive and requires a significant volume of labelled data. Furthermore, because of their image as "black boxes", it can be difficult to understand how CNNs make their decisions (Gu et al., 2018). Consequently, attempts have been made to provide techniques for assessing and comprehending these processes.

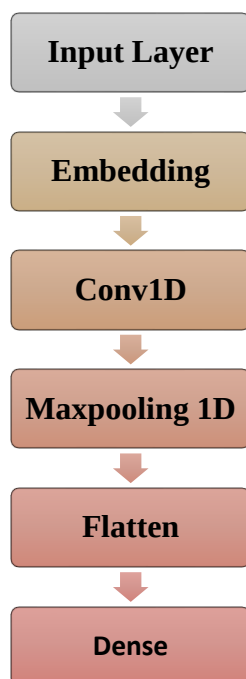
CNNs are algebraic equations that stimulate neural networks in the brain. The input, output, and hidden layers are the three basic levels that can be used by some systems, whereas others can only use the input and output layers. Several studies have looked into NLP in conjunction with sentiment analysis in the past few years, and the findings have been promising. The single model was outperformed by the CNN as well as LSTM strategies. Naive Bayes, Support Vector Machines, and Maximum Entropy are examples of other approaches whose results CNN has surpassed in many applications. CNN deployed a single convolutional layer to represent the grammar as a collection of words (Albawi et al., 2017).

CNNs can pick up several traits from the data they receive. They begin by grasping basic aspects like textures and edges at the levels, moving on to more intricate details like sizes and forms at higher layers. CNNs are efficient in tasks comprising image and video recognition due to their hierarchical structure. However, training a network can be resource-

and time-intensive, and requires a significant amount of tagged data. It can also be difficult to comprehend CNN's decision-making processes due to their perceived status as "black boxes" (Choubey, 2020). Consequently, attempts have been made to generate techniques for properly assessing and deciphering their predictions. The architecture of CNN is shown in Figure 8.

Figure 8

The CNN Structure



It should be mentioned, however, that CNN training demands enormous quantities of labelled data and can be resource- and time-consuming. Furthermore, understanding the logic underlying CNN identification may be difficult due to its classification as "black boxes" (Hema and Garcia Marquez, 2023). As such, efforts have been undertaken to offer techniques for the analysis and visualisation of the patterns that CNNs learn. In Figure 10, the CNN architecture is shown.

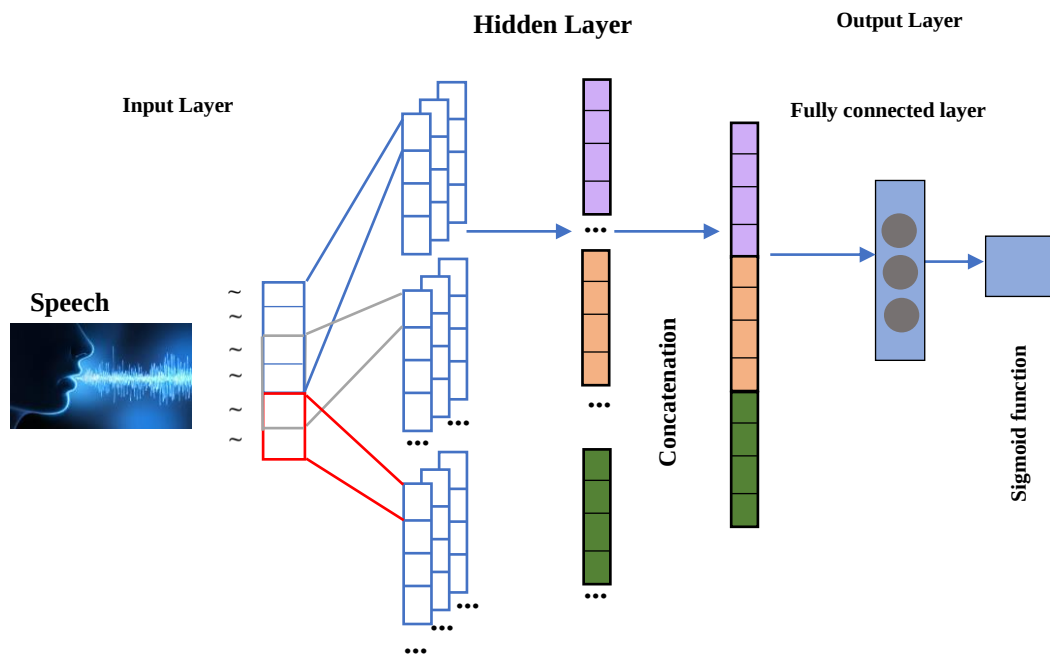
Many hidden layers are arranged among the input levels and output stages of a CNN (Al-Malla et al., 2022):

Convolution: Convolution employs a series of filters to process the input images.

Each filter recognises a unique feature and turns it on or off in response to it.

Figure 9

The CNN Architecture



Rectified Linear Units (ReLU): ReLUs provide for faster and more efficient training by maintaining positive values and setting negative values to zero. Simply said, "activation" refers to the transfer of only the active attributes to the following level.

Pooling Layer: Pooling uses non-linear downsampling to simplify the output and mitigate the total quantity of factors the network has to learn.

CNNs have become powerful tools in many AI applications, including ASR.

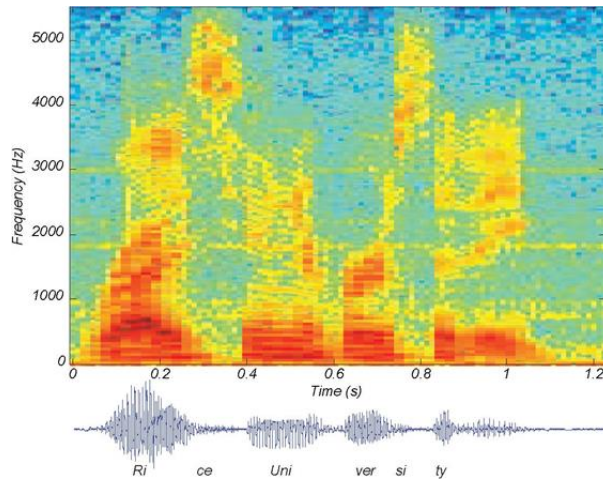
- **Feature Extraction**

CNN is very good at learning hierarchical data depiction on its own. The spectrogram of an acoustic wave, that presents the distribution of frequencies through time, is viewed as a picture in the field of ASR (Hema and Garcia Marquez, 2023). CNN is capable of extracting relevant characteristics from spectrograms in an efficient manner, which allows them to

capture important patterns that are essential for speech comprehension. A spoken signal's spectrogram is seen in Figure 10.

Figure 10

The Spectrogram Signal



- **Temporal Dependencies**

CNNs are mostly designed for spatial hierarchies, but they can also be used to manage the necessary temporal dependencies in sound identification through methods like 1D convolutions or CNNs combined with Recurrent Neural Networks (RNNs) (Choubey, 2020). Voice signals can be accurately represented by CNNs in terms of both the frequency content and temporal patterns.

Through the extraction of consistent sound information representations, CNNs develop the robustness of ASR systems. These systems are very useful for real-world-use situations since they can handle pronunciation, accent, and background noise variations.

CNNs are mostly designed for spatial hierarchies, but they can also be used to manage the necessary temporal dependencies in sound identification through methods like 1D convolutions, or CNNs combined with RNNs voice signals can be accurately represented by CNNs in terms of both the frequency content and temporal patterns through the extraction of consistent sound information representations (CNNs) to develop the robustness of ASR

systems (Choubey, 2020). These systems are very useful for real-world-use situations since they can handle pronunciation accents and background noise variations.

3.7.2 Bidirectional Long Short-Term Memory

The acronym for Long Short-Term Memory basically is LSTM. The LSTM network and its unit cell must first be understood to comprehend how Bidirectional Long Short-Term Memory (BI-LSTM) functions. According to their article (Hochreiter and Schmidhuber, 1997). The most often utilised neural networks are recurrent ones. The enhanced ability of RNNs to process sequence input is well known. Nonetheless, there are times when it's also required to preserve the outcomes of earlier data.

- **BI-LSTM**

Two LSTM layers make up a BI-LSTM, which is a type of series method. Whereas the other LSTM layer processes the input backwards, the first layer processes it forwards. NLP-related professions frequently use it. The rationale beyond this approach lies in the potential for the enhanced comprehension of sequential data by analysing the input data. This method aims to deepen the understanding of the interdependencies within a series, such as grasping the context provided by words preceding and following a given phrase (Schuster and Paliwal, 1997). Bidirectional LSTM systems function by receiving the training series in both forward and backward directions from two diverse LSTM systems that are attached to the same output level.

To be clear, both forward and backward orientations are used to encode the sequence. For every time stride, this entails concatenating the output from the front and backward LSTM. Words before and following it are now understood by the encoded representation of that word. To process the sequence in both the forward and backward directions, two unidirectional LSTMs are combined to form the bidirectional LSTM architecture. With two distinct LSTM networks, each with a distinct purpose, this design combines them. The token

sequence is processed by one LSTM network in the series order that it was originally intended, and by another LSTM network in the opposite order (Zhang et al., 2015). Both of the probability vectors that are generated as output are added at the end of each of these LSTM networks.

$$p_t = p_t^f + p_t^b$$

(1)

where p_t : is the final probability vector of the network

p_t^f : The probability vector of the front LSTM system

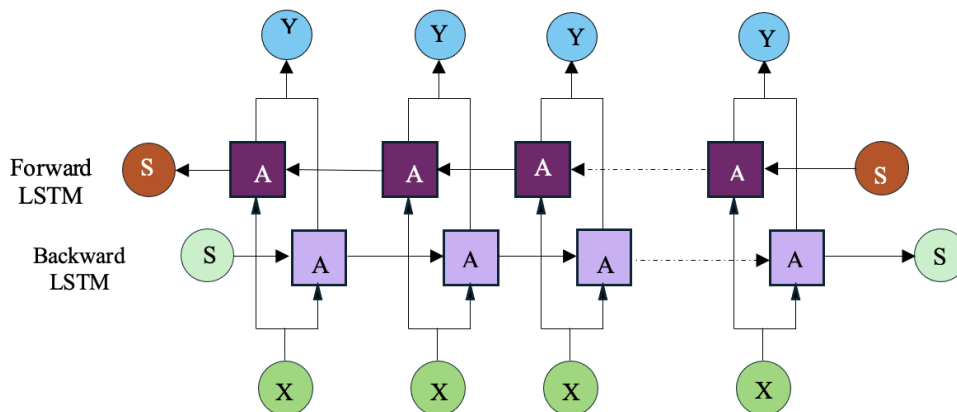
p_t^b : The probability vector of the rearwards LSTM system

Figure 11 illustrates the fundamental architecture of the BI-LSTM, where the LSTM nodes A and A', input token X_i , and output token Y_i are represented. Both the A and A' LSTM nodes make up Y_i 's final output (Zerari et al., 2019).

Next, we will explore building a review scheme in Python that leverages the BI-LSTM layers, utilising the TensorFlow framework. Initially, we will establish the system and train it to specify whether the feedback expresses negative, positive, or neutral sentiments.

Figure 11

The BI-LSTM Architecture



BI-LSTM systems have shown noticeable effectiveness in tasks involving sequences, making them highly suitable for ASR, the process of converting verbal linguistics into scripted text. This efficacy stems from their capabilities to capture both preceding and subsequent context, advancing the function of ASR systems (Zerari et al., 2019). As an extension of traditional LSTMs, BI-LSTMs excel in capturing long-term dependencies within series data.

The BI-LSTM algorithm comprises two LSTM layers: one processes the input sequence in the front direction, while the other processes it in reverse. This bidirectional architecture enables the network to capture dependencies in both forward and backward directions, leading to a more extensive conception of the series data (Zhang et al., 2015).

In the domain of ASR, BI-LSTM networks have become a foundational element, as they adeptly gather contextual data, thereby enhancing the model's ability to transcribe spoken English accurately. The bidirectional flow of data in BI-LSTM proves invaluable for representing the intricate patterns of speech accurately.

3.7.3 Gated Recurrent Unit

LSTM networks can be substituted with GRUs. This particular kind of RNN was first described by (Gao and Glowacka, 2016). As LSTM, the GRU can handle textual, acoustic, and sequential information, including time-series data.

The basis for the GRU's operation is the elective alteration of the network's hidden level through the gating technique at each time step. To control the stream of data entering and exiting the network, gating mechanisms are used. Two gating mechanisms make up the GRU: the reset gate and the update gate (Irie et al., 2016).

Another kind of RNS with memory cells is called a GRU. They have a simpler cell design than LSTM, although they are similar. Although it lacks of an output gate and has

fewer variables, the GRU likewise utilises a gating technique to organise the data stream through cell state.

A particular GRU single-cell structure is depicted in Figure 15. It has two gates: an update gate (z) and a reset gate (r). The update gate chooses the amount of prior memory to retain; meanwhile, the reset gateway guides the flow of new input into the preceding memory. The prior hidden state (h in Figure 12) is immediately attached to the reset gate when comparing the GRU with LSTM. The update gate is made up of the entrance and forget gates. The way the memory data are exposed differs as well. Because GRUs lack an exit gate, all of their memory data are visible; however, in LSTM, the memory data can be accessed or used by other units or cells in the output gate that controls the network (Chung et al., 2014).

The GRU output computations make use of the following equations:

$$\text{Reset gate:} \quad r_t = \text{sigm}(W_{xr} x_t + W_{hr} h_{t-1} + b_r) \quad (2)$$

$$\text{Update gate:} \quad z_t = \text{sigm}(W_{xz} x_t + W_{hz} h_{t-1} + b_z) \quad (3)$$

$$\text{Candidate hidden state:} \quad \tilde{h}_t = \tanh(W_{xh} x_t + W_{hh}(r_t \odot h_{t-1}) + b_h) \quad (4)$$

$$\text{Hidden state:} \quad h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \quad (5)$$

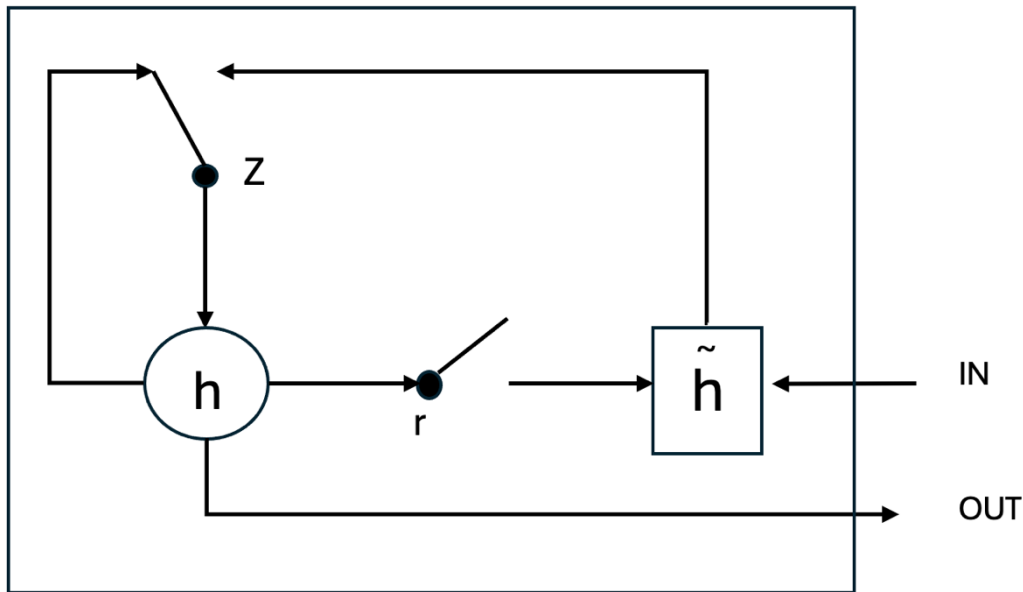
In Eqs. 2 through 5, x_t , h_t , r_t , and z_t stand for the input, output, reset, and update gates, correspondingly. The W parameters all represent weighting matrices, whereas the b variables represent biases. The hyperbolic tangent ($\tanh$) function, sigmoid function (sigm), and LSTM activation functions are all employed.

Both GRU and LSTM cells have a gating mechanism that makes them ideal for long-term dependence. Both of these cells were tested by (Bahdanau et al., 2016) for the machine translation task. Their initial testing demonstrated the comparability of the GRU and LSTM. Using datasets such as raw speech signal data and polyphonic music data, (Chung et al., 2014) evaluated LSTM and GRU cells in the sequence modelling task. LSTM and GRU networks

both performed well, but it was impossible to determine which one was superior. These trials gave me the idea to combine GRU and LSTM in this study's performance evaluation for the voice recognition problem.

Figure 12

The Structure of a Single Cell of the GRU



3.7.4 The Hybrid Convolution Neural Network-Bidirectional Long Short-Term Memory Approach

The CNN-BI-LSTM mechanism makes use of both CNN's capacity for spatial feature extraction and the temporal dependencies of BI-LSTM (Mishra et al., 2023). As CNN inputs into BI-LSTM, parsed phrases and documents are converted into numerical information that neural networks can understand. CNN-BI-LSTM models combine spatial feature representations by scanning input data for convolutional feature extraction and improving sequential modelling accuracy using one-dimensional filters or kernels. CNNs and BI-LSTMs handle sequential input and extract features by performing convolutional operations across sequential dimensions. Convolutional layer feature maps are reduced by the pooling

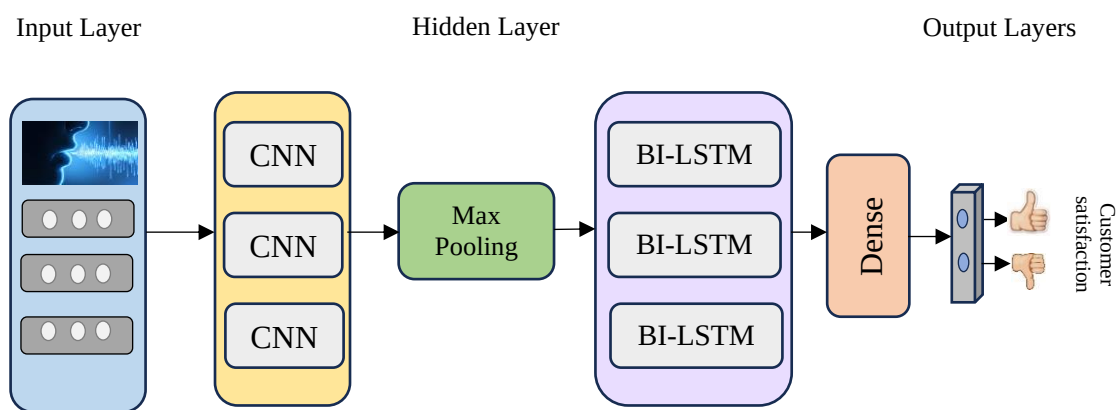
layer, and temporal and spatial representations are further refined by fully linked layers. Because of their adaptability and ability to extract spatial data from input data, CNN-BI-LSTM architectures are a good fit for tasks including text classification, sentiment analysis, and video analysis.

The CNN component, which is renowned for its capacity to extract spatial characteristics, analyses voice sound spectrogram representations concurrently. CNNs are able to capture particular patterns within a given locality, which enhances the model's capacity to discern relevant audio features (Mishra et al., 2023). When these two architectures are combined, a hybrid model is produced that is very good at learning intricate temporal and spatial representations, which are necessary for ASR. The CNN-BI-LSTM architecture for ASR is shown in Figure 13.

- **CNN-BI-LSTM Performance in ASR**

Figure 13

The CNN-BI-LSTM Architecture



The CNN-BI-LSTM model requires multiple stages to operate in ASR. First, the model receives acoustic waveforms and transforms them into representations of spectrograms. After that, the CNN component extracts spatial features, highlighting significant patterns in the spectrogram. The BI-LSTM component simultaneously records

contextual data and long-term dependencies. The final layers of the algorithm effectively convert verbal linguistics into text by merging these collected attributes with a forecast of the matching transcription (Mishra et al., 2023).

Because it is more accurate and resilient than traditional models, the CNN-BI-LSTM hybrid model is crucial for ASR applications. The intricate nature of spoken language aligns with the ability to effectively capture both spatial and time-based properties. The aforementioned characteristic makes it extremely advantageous in an extensive array of uses, including voice assistants, transcription services, and speech-activated devices (Mishra et al., 2023).

In summary, a major advancement in the field of automatic speech recognition is the CNN-BI-LSTM hybrid model. This model shows how well bidirectional long short-term memory and Convolutional Neural Network architectures may be combined to meet spoken language issues. As voice recognition technology advances, the CNN-BI-LSTM model offers more accuracy and versatility.

3.7.5 Hybrid Gated Recurrent Units -Bidirectional -Long Short-Term Memory

In the realm of ASR, the Gated Recurrent Units -Bidirectional -Long Short-Term Memory GRU-BI-LSTM algorithm is a novel design that combines the benefits of BI-LSTM systems and GRUs. This study investigates the design, functionality, and important impact of the GRU-BI-LSTM model on developing the accuracy and functionality of the ASR method (Vashisht et al., 2024).

- **The GRU-BI-LSTM Approach Structure**

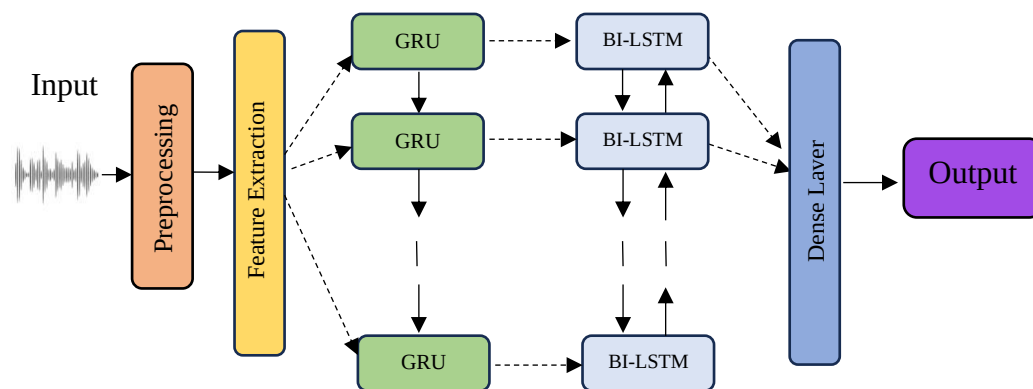
GRUs and BI-LSTMs are two powerful variants of RNNs combined in the GRU-BI-LSTM model. Each GRU unit has reset and update gates that determine what data should be updated and what should be forgotten. Simultaneously, the forward and backward orientations of input sequences are analysed by the BI-LSTM, an improved variant of the

traditional LSTM. The GRU-BI-LSTM architecture's hierarchical structure and connections are shown in Figure 14.

RNNs with memory cells also come in the form of GRUs. Although their cell design is simpler, they are comparable to the LSTM. In addition to having fewer variables and no output gate, the GRU also has a gating mechanism to regulate the data stream through the cell state.

Figure 14

The GRU-BI-LSTM Structure



The GRU-BI-LSTM model is highly skilled at ASR due to its capability of extracting contextual data and serial dependencies inside acoustic signals. The first stage in the ASR pipeline is to transform raw audio inputs into spectrogram representations. The input of the GRU-BI-LSTM model is these representations. The BI-LSTM evaluates the sequential context both forward and backwards, but the GRU units in the model are capable of selectively remembering or forgetting data. The model dynamically adjusts to variations in speech patterns, improving the transcription accuracy (Munir et al., 2021).

3.8 Evaluation Report

Measures including the accuracy, precision, recall, and F1 score are frequently applied to assess a classification mechanism's efficacy. According to (Naidu et al., 2023), these metrics are categorisation-pertinent for assessing models that forecast binary results,

such as identifying between spam and non-spam, fraud and non-fraud, or positive and negative.

Accuracy is a fundamental and often used measure that calculates the ratio of accurately predicted occurrences to the whole dataset. It functions as a metric of the model's ability to accurately classify occurrences into their assigned categories. Generally speaking, the accuracy statistic counts the proportion of accurate forecasts to all instances assessed (Naidu et al., 2023).

The accuracy is calculated using Eq. 6:

Accuracy = (true positives + true negatives) / (true positives + false positives + true negatives + false negatives)

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \quad (6)$$

Precision is a formula that assesses the percentage of accurately predicted positive situations in all cases that the mechanism predicts to be positive (Naidu et al., 2023). The term "signifies" refers to the model's predictive accuracy in positive circumstances, that is ascertained by using a particular application of Eq. 7:

Precision = true positives / (true positives + false positives)

$$Precision (P) = \frac{TP}{TP+FP} \quad (7)$$

Recall is an alternative scale called recall that counts the percentage of correctly recognised positive conditions relative to all positive examples in the collection. The metric, which is calculated using Eq. 8, assesses how well the model detects positive cases (Naidu et al., 2023).

Recall = true positives / (true positives + false negatives)

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

(8)

The **f1 score** achieves an equilibrium amidst the two metrics and is computed as the harmonic mean of accuracy and recall. Recall and accuracy are both very significant (Naidu et al., 2023). Eq. 9 computes the f1 score:

$$F1 \text{ score} = 2 \text{ (precision recall) / (precision + recall)}$$

$$F1 - score = \frac{2*P*R}{P+R} \quad (9)$$

The Confusion Matrix is a computational statistical evaluator where the accuracy of classification techniques could be assessed using known as a confusion matrix, which is made up of false positive, true positive, false negative, and true negative. In order to assess the accuracy and performance of a machine-learning classifier on a multi-label classification dataset, the confusion matrix is additionally represented in a table that might be made (Gaye and Wulamu, 2019).

Key Components:

1. True Positives (TP): Situations where the model accurately predicts the Positive class.
2. True Negatives (TN): Situations where the model accurately predicts the absence of the Positive class.
3. False Positives (FP): Also known as Type I mistakes, occur when the model predicts the Positive class incorrectly.
4. False Negatives (FN): Also known as Type II mistakes, are cases where the model predicts the Negative class incorrectly.

Binary classification: Table 7 illustrates the configuration of the confusion matrix for polarity classification (Gaye and Wulamu, 2019).

Table 7*Confusion Matrix Polarity Categorisation*

	Predicted positive	Predicted negative
Actual positive	True positives	False negatives
Actual negative	False positives	True negatives

3.9 Summary

This chapter outlines the techniques utilised in detail. It explains the steps taken to collect and analyse data, such as extracting features, dividing the data, and adding annotations. Moreover, various deep learning structures are discussed in this chapter, including networks CNNs BI-LSTM, and hybrid CNN-BI-LSTM approaches. The chapter wraps up with an explanation of the assessment criteria applied in the dissertation.

Chapter 4: Implementation and Experiment

4.1 Deployment Environment

Making the appropriate platform choice for the sound recognition system progression is mandatory due to the suggested mechanism's application calls for the use of a variety of tools and concepts. These are the libraries and tools used during implementation:

TensorFlow is considered a Google Brain-developed, open-source framework for machine learning that supports the creation and deployment of learning models on platforms. It comes with user APIs like Keras for easy model development and advanced APIs for added flexibility.

PyTorch, created by Facebook's AI Research Lab, is another source deep learning framework known for its dynamic computational graph and intuitive interface. It is widely used in research and production environments, offering an effective platform for constructing networks.

Keras, originally an independent project, Keras has now been integrated into TensorFlow as its high-level API. Keras simplifies the process of building networks with coding and is commonly used for quick prototyping and experimentation.

Scikit learns, is primarily recognised for machine learning algorithms, it also includes tools for network models such as multi-layer perceptron and deep belief networks. It is a choice for tasks like data preparation, model assessment, and ensemble learning.

Librosa is a Python package for music and audio analysis. It provides a wide range of functionalities for tasks such as loading audio files, extracting features from audio signals, and performing various audio processing tasks. Librosa is particularly useful for tasks related to audio signal processing, music information retrieval, and machine learning on audio data.

Google Colab (short for Google Colaboratory) is a cloud-based platform provided by Google that offers free access to GPU and TPU resources for running Python code. It is built

on top of Jupyter Notebooks and allows users to write and execute Python code in a web-based environment. Colab is widely used for various data science and machine learning tasks, as it provides a convenient way to run code, visualise data, and collaborate with others without requiring any setup or installation.

The Colab platform runs on an Ubuntu 17.10 64-bit system, powered via an undisclosed Intel Xeon processor with two cores operating at 2.3 GHz. It boasts a total of 13 GB RAM. Additionally, it is equipped with an NVIDIA Tesla K80, housing a GK210 chipset, 8 GB of RAM, and 2496 CUDA cores clocked at 560 MHz (Carneiro et al., 2018). The proposed model will be executed on Colab, leveraging its cost-free Jupyter Notebook interface.

4.2 Experimental Architecture

4.2.1 *Gathering and Preparing the Dataset*

The Arabic voice emotion recognition dataset's integrity and trustworthiness were ensured by the study's accurate data collection and formatting process. This section provides a brief synopsis of the procedures used to gather and prepare the data.

- **Data Gathering**

The method of obtaining Arabic audio comments for COP28 involved gathering information from multiple websites, including YouTube and Instagram web pages. The procedure of obtaining reviews was methodical, with the goal of obtaining a wide range of reviews from these platforms. This technique guaranteed the inclusivity of the dataset by encompassing a wide range of sentiment analysis and emotional categories. (Please refer to section 3.1 for more details.)

- **Preprocessing of the Data**

One core step in preparing data for analytical and machine learning projects is data preparation. This procedure comprises transforming unprocessed, raw data into a format that

is standardised, structured, and refined – perfect for efficient algorithmic use. This initial phase greatly improves the quality and accuracy of the analytical outcomes. After the raw data were gathered, a series of preprocessing steps were taken to prepare it for investigation and classification, as explained in section 3.3.

4.2.2 Models Utilised in Arabic Speech Recognition

Strong DL structures and CNNs are frequently implemented for applications like acoustic classification. A CNN's functionality is greatly impacted by its hyperparameters, which are configurations that specify the design and operation of the system. The hyperparameters used in this investigation except for the CNN mechanism in the ASR with COP28 as a special case study dataset are displayed in Table 8 which follows.

- **CNN**

There are several hyperparameters used in building and compiling the Convolutional Neural Network CNN model. Hyperparameters are parameters whose values are set before the training process begins and influence the model's architecture and behaviour. Table 8 below demonstrates the hyperparameters in CNN.

Table 8

The Hyperparameter Setup for the CNN Model

Layer	Hyperparameter	Value
Convolutional Layer	Number of filters	128
	Kernel size	3
	Activation function	ReLU
	Padding	'same'
MaxPooling1D Layer	This layer performs a global max pooling operation over the entire sequence	
	pool_size	2
Dense Layer (Hidden Layer)	Number of units	128

	Activation function	ReLU
Dropout Layer	Dropout rate	0.2
Dense Layer (Hidden Layer)	Number of units	128
	Activation function	ReLU
Dense Layer (Output Layer)	Number of units	3
	Activation function	Softmax
	epochs	100
	batch_size	512

Below is a break down for each layer in the provided model:

1. Conv1D Layer:

- This layer performs one-dimensional convolution over the input sequence.
- `Conv1D(128, kernel_size=3, activation='ReLU', input_shape=(X.shape[1], 1))`
- It has 128 filters, each of size 3.
- The activation function used is ReLU.
- The input shape is `(X.shape[1], 1)`, where `X.shape[1]` represents the length of the input sequence, and `1` indicates the input dimensionality (1 channel in this case).

2. MaxPooling1D Layer:

- This layer performs one-dimensional max pooling over the output of the convolutional layer.
- `MaxPooling1D(pool_size=2)`
- It reduces the spatial dimensions of the input by taking the maximum value within a window of size 2.

3. Flatten Layer:

- This layer flattens the output of the previous layer into a one-dimensional array.
- `Flatten()`
- It prepares the data for input into the fully connected layers by converting the multi-dimensional output into a flat vector.

4. Dense Layers:

- These are fully connected layers that perform operations on the flattened input.
- ``Dense(128, activation= ReLU)`` followed by ``Dense(128, activation= ReLU)``
- Both layers have 128 neurons and use the ReLU activation function.
- They allow the network to learn complex patterns from the flattened input data.

5. Dropout Layers:

- These layers introduce dropout regularisation to prevent overfitting.
- ``Dropout(0.2)`` followed by ``Dropout(0.2)``
- A dropout rate of 0.2 means that 20% of the input units will be randomly set to zero during training, helping to prevent the network from relying too much on specific activations and improving its generalisation ability.

6. Output Layer (Dense):

- This layer produces the final output of the network.
- ``Dense(3, activation='softmax')``
- It has 3 neurons, corresponding to the number of output classes in the classification task.
- The activation function used is softmax, which converts the output scores into probabilities, making it suitable for multi-class classification problems.

These layers collectively define the architecture of the CNN model, from processing the input data to producing the final predictions. Each layer performs specific operations that contribute to the overall functionality and learning capabilities of the network.

• **BI-LSTM**

In this dissertation, a BI-LSTM model is identified utilising the sequential API of the Keras. Table 9's model architecture and hyperparameters are explained as follows:

Table 9*The Hyperparameters' Setup of the BI-LSTM*

Layer	Hyperparameter	Value
BI-LSTM Layer	The number of utilised units	40
	Return sequence	True (for the first BI-LSTM layer)
Dense Layer (Hidden Layer)	The number of used units	123
	The type of activation function	ReLU
Dropout Layer	The number of the dropout rate	0.2
Dense Layer (Hidden Layer)	The number of used units	32
	The type of activation function utilised	ReLU
Dense Layer (Output Layer)	The number of applied units	3
	The type of activation function applied	Softmax
	The number of epochs	100
	The size of the applied batch	512

The Model Architecture:

1. BI- LSTM Layer:

- This layer contains a Bidirectional Long Short-Term Memory (LSTM) unit.
- `(BI-LSTM(123, input\_shape=(40, 1)))`
- The LSTM layer has 123 units.
- It processes sequences bidirectionally, allowing the model to capture dependencies in both forward and backward directions.

- The input shape is `(40, 1)`, indicating a sequence length of 40 and an input dimensionality of 1.

2. Dense Layers:

- Two dense layers follow the Bidirectional LSTM layer.

- `Dense(128, activation= ReLU)` followed by Dense(32, activation= ReLU)``
- The first dense layer has 128 neurons, and the second dense layer has 32 neurons.
- Both layers use the ReLU activation function to introduce non-linearity into the model.

3. Dropout Layers:

- Two dropout layers are used for regularisation to prevent overfitting.
- `Dropout(0.2)` followed by Dropout(0.2)``
- A dropout rate of 0.2 means that 20% of the input units will be randomly set to zero during training.

4. Output Layer (Dense):

- The final dense layer produces the output of the model.
- `Dense(6, activation='softmax')``
- It has 6 neurons, representing the number of output classes in the classification task.
- The activation function used is softmax, which converts the output scores into probabilities.

The Hyperparameters Setup:

- Loss Function:
 - The model is compiled with a 'categorical_crossentropy' loss function, suitable for multi-class classification tasks where there are more than two classes.
- Optimiser:
 - The 'Adam' optimiser is used for optimising the model's weights during training.

Adam is an adaptive learning rate optimisation algorithm that is well-suited for most deep learning tasks.
- Metrics:

- The model is evaluated based on 'accuracy' during training.
- Training Parameters:
 - `validation\_split=0.2`: 20% of the training data is used as validation data for monitoring the model's performance during training.
 - `epochs=100`: The model is trained for 100 epochs, iterating over the entire dataset 100 times.
 - `batch\_size=512`: The number of samples per gradient update. The training data are divided into batches of size 512 for each training iteration.
 - `shuffle=True`: The training data are shuffled before each epoch to introduce randomness and prevent the model from memorising the order of the training samples.

Training and Evaluation:

- Training:
 - The model is trained using the `fit` method on the training data `X` and one-hot encoded label `y\_one\_hot`.
- Evaluation:
 - After training, the model's predictions are evaluated by computing the predicted classes and comparing them against the true classes to calculate the accuracy.

• GRU

A certain sort of RNN design known as the GRU seeks to solve the vanishing gradient problem, which is one of the limitations of traditional RNNs. GRUs were developed (Chung et al., 2014) and used gating approaches to reset and refresh data in a selective manner. They are especially good at capturing long-term associations in serial data because of this property.

The hyperparameters implemented in the GRU approach for the Arabic speech recognition datasets are displayed in Table 10 below.

Table 10*GRU Hyperparameters*

Hyperparameters	Value
The number of GRU units	225
The utilised activation function	tanh
The utilised dropout rate	0.2
The applied optimiser	adam
The implemented learning rate	0.001
The utilised batch size	512
Number of epochs	100
The utilised loss function	sparse_categorical_crossentropy

A GRU model is built and compiled for a classification task. Let us break down the hyperparameters used in this model:

1. GRU Layer:

- Units: 256 - This specifies the dimensionality of the output space (i.e., the number of units in the GRU layer).

- Input shape: `(X.shape[1], 1)` - This defines the shape of the input data. Here, `X.shape[1]` represents the number of time steps in the input sequence, and `1` signifies that each time step has a single feature.

2. Dense Layers:

- Two dense layers with 128 units each and ReLU activation are added after the GRU layer. These layers help in learning complex patterns in the data.

3. Dropout Layers:

- Dropout layers with a dropout rate of 0.2 are inserted after each dense layer.

Dropout is a regularisation technique used to prevent overfitting by randomly dropping a fraction of units (0.2 in this case) during training.

4. Output Layer:

- A dense layer with 3 units and softmax activation is added as the output layer.

This layer produces the final output probabilities for each class.

5. Compilation:

- Loss function: Categorical Crossentropy - This is the loss function used to compute the model's loss during training. It is suitable for multi-class classification tasks.
- Optimiser: Adam - The Adam optimiser is used to optimise the model's weights during training.
- Metrics: Accuracy - Accuracy is chosen as the metric to evaluate the performance of the model during training.

6. Training:

- Training data (`X`, `y\_one\_hot`) is used to train the model.
- Validation split: 0.2–20% of the training data are used as validation data to monitor the model's performance during training.
- Epochs: 100 - The number of epochs determines how many times the entire training dataset is passed forward and backwards through the neural network.
- Batch size: 512 - The number of samples processed before the model's weights are updated.

Overall, these hyperparameters configure the architecture and training process of the GRU model for the given classification task. Adjusting these hyperparameters can have a significant impact on the model's performance and training dynamics.

- **CNN- BI-LSTM**

CNN-BI-LSTM is a powerful DL architecture often used for image and text categorisation problems. A CNN's performance is significantly influenced by its hyperparameters, which are configurations that determine the structure and operation of the network. The hyperparameters implemented in the CNN-BI-LSTM approach for the Arabic speech recognition datasets are displayed in Table 11 below.

Table 11

The CNN-BI-LSTM Hyperparameters

Layer	Hyper-parameter	Value
Convolutional Layer	The number of applied filters	128
	The kernel size implemented	3
	The type of activation function utilised	ReLU
	The type of padding	'same'
GlobalMaxPooling1D Layer	The pool size applied	3
Bidirectional LSTM Layers	The number of units implemented	123
	Return sequences	True
Dense Layer (Hidden Layer)	The number of units applied	128
	The type of activation function applied	ReLU
Dropout Layer	Dropout rate	0.2
Flatten Layer	Number of units	256
	Activation function	ReLU
Dense Layer (Output Layer)	Number of units	3
	Activation function	Softmax
	The number of epochs	100
	batch_size	512

This model represents the construction and training of a CNN followed by a BI-LSTM model using Keras with a TensorFlow backend. Let us break down the components and hyperparameters involved:

1. CNN Layers:

- Conv1D: A one-dimensional convolutional layer with 128 filters, a kernel size of 3, and ReLU activation function. It takes input sequences of shape (`sequence_length`, 1).

- MaxPooling1D: A max-pooling layer with a pool size of 3.

2. BI-LSTM Layer:

- Bidirectional: This wrapper creates a BI- LSTM layer with 123 units. Bidirectional LSTM processes the input sequence in both forward and backward directions, which can capture information from both past and future contexts.

3. Dense Layers:

- Dense: A fully connected layer with 128 units and ReLU activation function.

- Dropout: A regularisation technique that randomly drops a fraction of input units during training to prevent overfitting. In this case, the dropout rate is set to 0.2.

- Another dense layer with 256 units and ReLU activation followed by a dropout layer with a dropout rate of 0.2.

- Final dense layer with 3 units and softmax activation for multi-class classification.

This layer outputs the probabilities of each class.

4. Custom Learning Rate:

- The custom learning rate is set to 0.0001.

5. Compilation:

- Loss Function: Categorical crossentropy is chosen as the loss function since it is commonly used for multi-class classification tasks.

- Optimiser: The Adam optimiser is used with default parameters.

- Metrics: Accuracy is used as the evaluation metric.

6. Training:

- Training Data: `X` represents the input data, and `y\_one\_hot` represents the one-hot encoded target labels.
- Validation Split: 20% of the training data are used as validation data.
- Epochs: The model is trained for 100 epochs.
- Batch Size: Training is performed in mini-batches of size 512.
- Shuffle: The training data are shuffled before each epoch.

After training, the training history is stored in the `history` variable, which can be used to visualise training and validation metrics like accuracy over epochs.

- **GRU- BI-LSTM**

BI-LSTM layers and GRU layers are integrated into the GRU-BI-LSTM hybrid approach. Capturing serial links and patterns in data is the aim of this design. The BI-LSTM and GRU layers function together to extract and represent short-term and long-term relationships from serialised data.

Table 12 illustrates the hyperparameters of the GRU-BI-LSTM hybrid model.

Table 12

The Hyperparameters of the GRU-BI-LSTM Hybrid Model

Hyperparameters		Value
Number of units in the GRU layer	gru_units	64
Number of units in the BI-LSTM layer	lstm_units	64
Data about the convolutional layers	Conv_Filters	64
	convolutional kernel size	(3, 3)

	activation function	ReLU
	pool size	(2, 2)
Data about the dense layer	number of LSTM units	64
	dense_units	256
	dropout	0.2
Type of utilised optimiser	optimiser	Adam
	learning rate	0.0001
	batch size	512
	epochs	100

The GRU-BI-LSTM hybrid approach has the following hyperparameters that are described as follows:

1. GRU Units:

- The first layer is a GRU layer with 64 units.
- GRU is a type of RNN that is effective in capturing long-range dependencies in sequential data.

2. Return Sequences:

- ``return_sequences=True`` indicates that the GRU layer should return sequences instead of just the last output.

3. Input Shape:

- ``input_shape=(X.shape[1], 1)`` specifies the shape of the input data.
- ``X.shape[1]`` represents the number of time steps in the input sequence, and ``1`` indicates that each time step has only one feature.

4. Bidirectional LSTM:

- The output of the GRU layer is passed to a BI- LSTM layer with 123 units.
- Bidirectional LSTM processes the input sequence in both forward and backward directions, which can capture information from both past and future contexts.

5. Dense Layers:

- Two dense layers with 128 and 64 units, respectively, followed by ReLU activation functions.
- ReLU is a popular activation function that introduces non-linearity to the model.

6. Dropout:

- Two dropout layers with a dropout rate of 0.2 are added after the dense layers.
- Dropout is a regularisation technique used to prevent overfitting by randomly dropping a fraction of input units during training.

7. Output Layer:

- The final dense layer consists of 3 units with a softmax activation function.
- Softmax activation is commonly used in multi-class classification problems to convert raw predictions into probabilities.

8. Compilation:

- The model is compiled with categorical crossentropy loss, which is suitable for multi-class classification tasks.
- The Adam optimiser is used for optimisation, which is an adaptive learning rate optimisation algorithm.

9. Metrics:

- The model is evaluated based on accuracy during training.

10. Training Parameters:

- ``validation_split=0.2`` specifies that 20% of the training data will be used for validation during training.
- ``epochs=100`` indicates the number of times the entire dataset will be passed forward and backwards through the neural network.

- `batch\_size=512` defines the number of samples that will be propagated through the network before updating the model's parameters.

- `shuffle=True` indicates that the training data will be shuffled before each epoch to introduce randomness and prevent the model from memorising the order of the data.

4.3 Results and Discussion

Investigating COP28 for Arabic speech recognition utilising a range of models including CNN, BI-LSTM, GRU, and hybrid combinations like CNN-BI-LSTM and GRU-BI-LSTM provides valuable insights into the performance and efficacy of these models within the distinct context of the field.

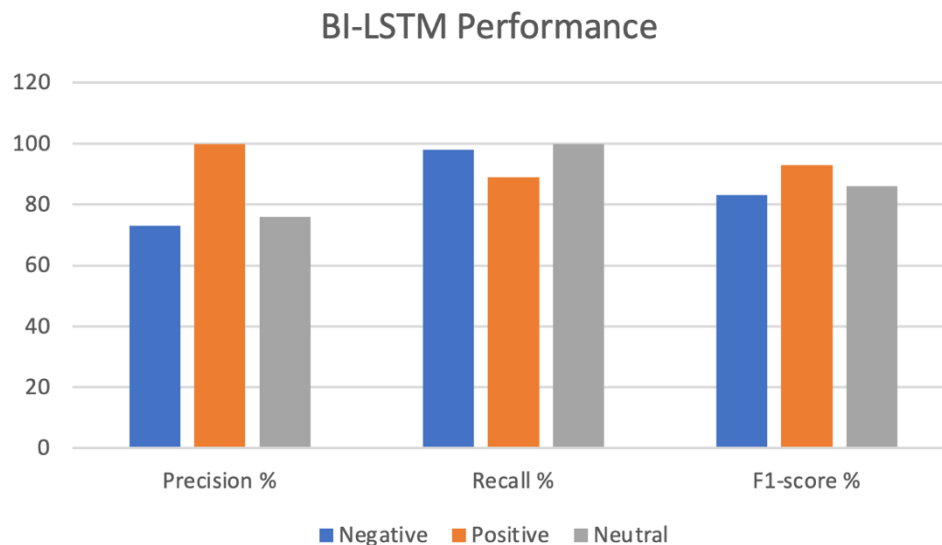
Table 11 and Figure 18 illustrate the BI-LSTM performance examining COP28 for Arabic speech recognition.

The table presents the evaluation results for the COP28 Reviews dataset across different classes, including Negative, Positive, and Neutral. Each class is assessed based on precision, recall, and F1-score metrics.

Table 13

The Performance of BI-LSTM

Dataset	Classes	Precision %	Recall %	F1-score %
COP28 Reviews	Negative	73	98	83
	Positive	100	89	93
	Neutral	76	100	86

Figure 15*The Performance of the BI-LSTM*

For the Negative class:

- Precision is 73%, indicating that among the instances predicted as negative, 73% were correctly classified.

- Recall is 98%, suggesting that among all the actual negative instances, 98% were correctly identified by the model.

- The F1-score is 83%, which is the harmonic mean of precision and recall, representing the balance between them.

For the Positive class:

- Precision is 100%, indicating that all instances predicted as positive were indeed positive.

- Recall is 89%, implying that 89% of the actual positive instances were correctly classified by the model.

- The F1-score is 93%, reflecting the overall performance in terms of precision and recall.

For the Neutral class:

- Precision is 76%, indicating that 76% of the instances predicted as neutral were indeed neutral.

- Recall is 100%, suggesting that all actual neutral instances were correctly identified by the model.

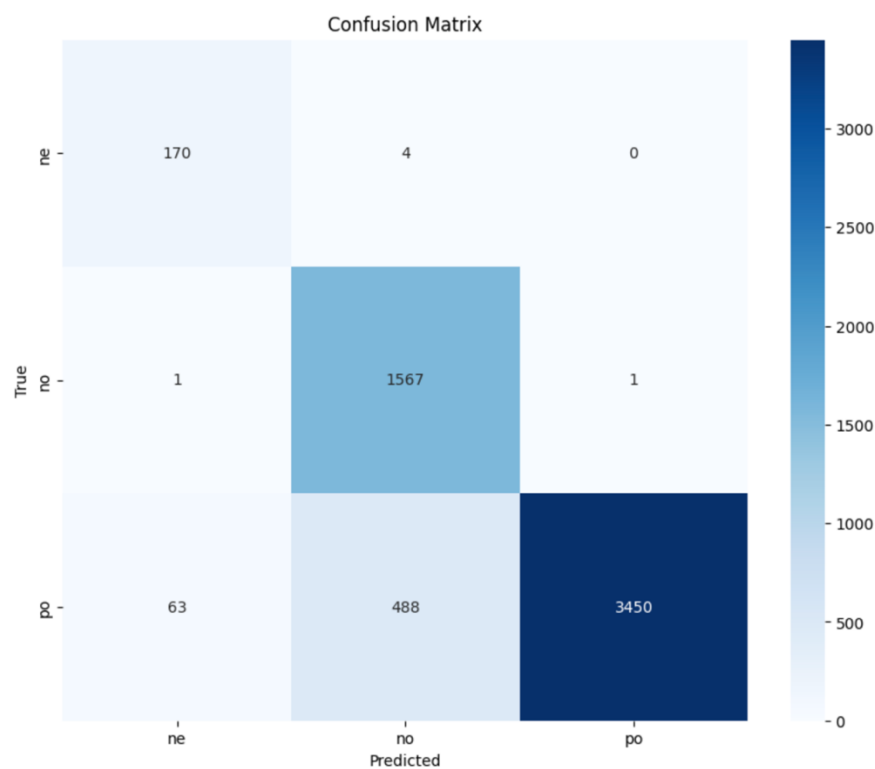
- The F1-score is 86%, representing the balance between precision and recall for the Neutral class.

Overall, these results provide insights into the performance of the model across different classes in the COP28 Reviews dataset, highlighting the strengths and areas for improvement in classification accuracy.

The confusion matrix in Figure 19 presents the classification results for the BI-LSTM performance.

Figure 16

The Confusion Matrix for the BI-LSTM



Here is an interpretation of the confusion matrix:

For the Negative class:

170 instances of the first class were correctly classified as the first class (True Negatives).

4 instances of the first class were misclassified as the second class (False Positives).

0 instances of the first class were misclassified as the third class (False Positives).

For the Neutral class:

1 instance of the second class was misclassified as the first class (False Negatives).

1,567 instances of the second class were correctly classified as the second class (True Positives).

1 instance of the second class was misclassified as the third class (False Positives).

For the Positive class:

63 instances of the third class were misclassified as the first class (False Negatives).

488 instances of the third class were misclassified as the second class (False Negatives).

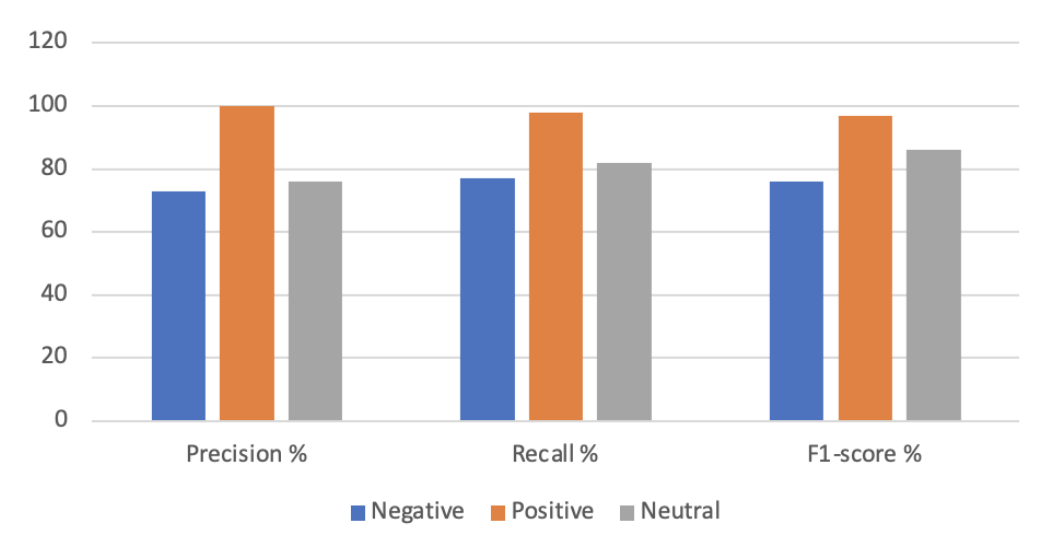
3,450 instances of the third class were correctly classified as the third class (True Positives).

Each row represents the actual class, while each column represents the predicted class. The diagonal elements represent correctly classified instances, while the off-diagonal elements represent misclassifications. This matrix provides a detailed overview of the model's performance across different classes, enabling further analysis and potential improvements.

Table 14 and Figure 17 illustrate the CNN performance examining COP28 for Arabic speech recognition.

Table 14*The CNN Performance*

Dataset	Classes	Precision %	Recall %	F1-score %
COP28 Reviews	Negative	82	99	90
	Positive	78	100	88
	Neutral	100	88	94

Figure 17*The CNN Performance*

The provided table summarises the evaluation results for the COP28 Reviews dataset across three distinct classes: Negative, Positive, and Neutral. Each class is assessed based on key metrics including precision, recall, and F1-score.

For the Negative class:

- The precision is 82%, indicating that among instances predicted as negative, 82% were correctly classified.

- The recall is 99%, signifying that the model correctly identified 99% of the actual negative instances.

- The F1-score stands at 90%, reflecting the balance between precision and recall for the Negative class.

For the Positive class:

- The precision is 78%, suggesting that 78% of the instances predicted as positive were indeed positive.

- The recall is 100%, indicating that all actual positive instances were correctly classified by the model.

- The F1-score is 88%, representing the overall performance in terms of precision and recall for the Positive class.

For the Neutral class:

- The precision is 100%, implying that all instances predicted as neutral were indeed neutral.

- The recall is 88%, indicating that 88% of the actual neutral instances were correctly identified by the model.

- The F1-score is 94%, highlighting the balance between precision and recall for the Neutral class.

These results offer valuable insights into the model's performance across different classes within the COP28 Reviews dataset, demonstrating its ability to effectively classify instances into their respective categories.

The confusion matrix in Figure 18 presents the classification results for the CNN performance.

From the figure, each row represents the actual class, while each column represents the predicted class. Below is an explanation of the results obtained from the confusion matrix:

For the Negative class, 172 reviews were correctly classified as negative, 2 were incorrectly classified as neutral, and none were classified as positive.

For the Neutral class, 1,568 samples were correctly classified as neutral, while 0 was incorrectly classified as negative, and 1 was incorrectly classified as positive.

For the Positive class, 3,426 samples were correctly classified as positive, while 37 were incorrectly classified as negative, and 438 were incorrectly classified as neutral.

Figure 18

The Confusion Matrix for the CNN Model

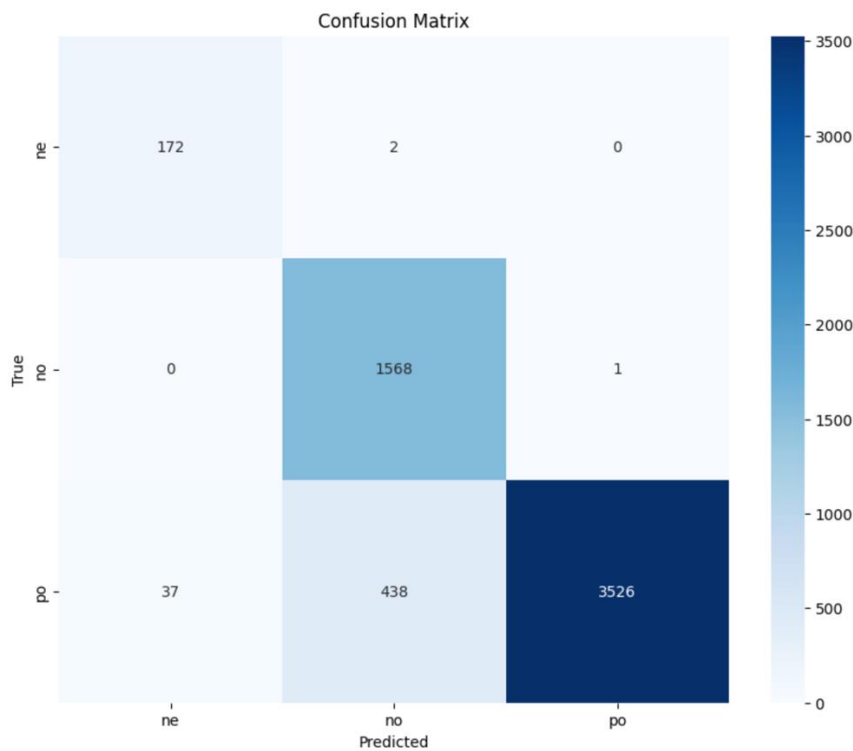


Table 15 and Figure 19 illustrate the GRU performance examining COP28 for Arabic speech recognition.

The provided table and figure present the evaluation metrics for the COP28 Reviews dataset across three distinct classes: Negative, Positive, and Neutral. These metrics, including precision, recall, and F1-score, offer valuable insights into the performance of the classification model.

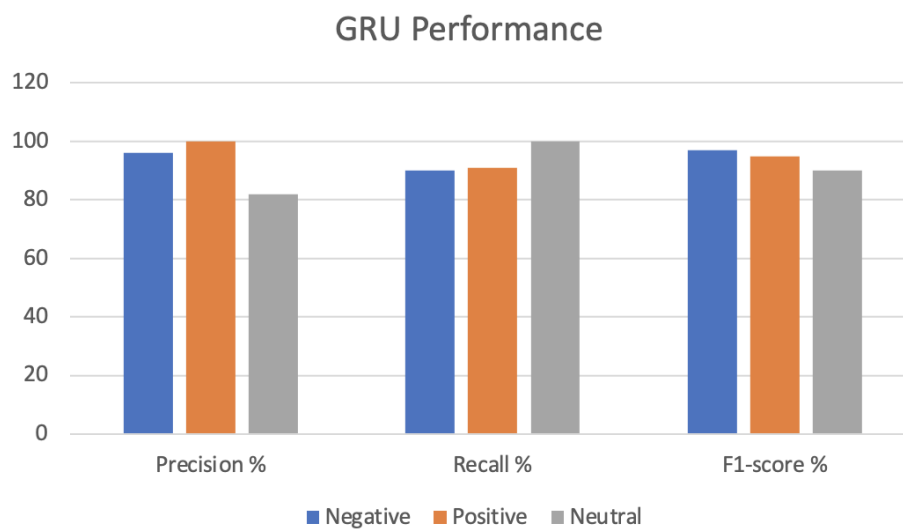
Table 15

The GRU Performance

Dataset	Classes	Precision %	Recall %	F1-score %
COP28 Reviews	Negative	96	90	97
	Positive	100	91	95
	Neutral	82	100	90

Figure 19

The GRU Performance



For the Negative class:

- Precision is 96%, indicating that among the instances predicted as negative, 96% were correctly classified.

- Recall is 90%, suggesting that the model correctly identified 90% of the actual negative instances.

- The F1-score, the harmonic mean of precision and recall, is 97%, representing a balanced performance in capturing both precision and recall for the Negative class.

For the Positive class:

- Precision is 100%, implying that all instances predicted as positive were indeed positive.

- Recall is 91%, indicating that 91% of the actual positive instances were correctly classified by the model.

- The F1-score is 95%, reflecting the overall effectiveness of the model in achieving a balance between precision and recall for the Positive class.

For the Neutral class:

- Precision is 82%, suggesting that 82% of the instances predicted as neutral were indeed neutral.

- Recall is 100%, demonstrating that the model correctly identified all actual neutral instances.

- The F1-score is 90%, highlighting the balance between precision and recall for the Neutral class.

These results provide a comprehensive overview of the model's performance across different classes within the COP28 Reviews dataset, showcasing its ability to accurately classify instances into their respective categories while maintaining a balance between precision and recall.

The provided confusion matrix as in Figure 20, offers a detailed representation of the classification performance of a model across multiple classes. Each row of the matrix corresponds to the actual class, while each column represents the predicted class.

Here's a breakdown of the confusion matrix:

- In the first row (for the Negative class):

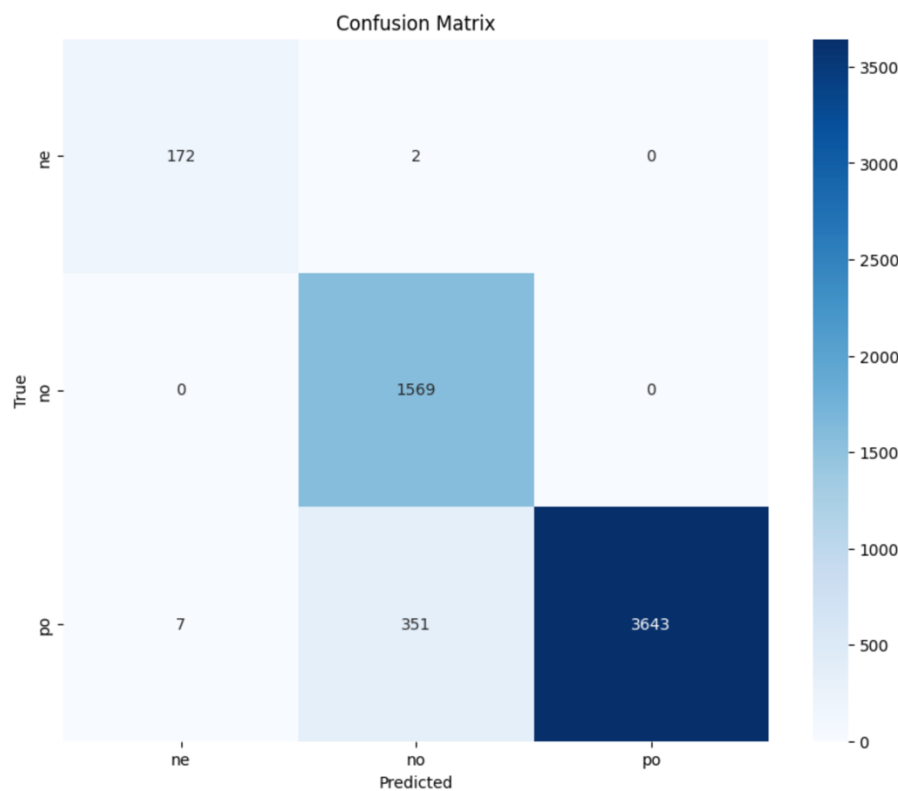
- 172 reviews of the first class were correctly classified as the first class (True Negatives).

- 2 reviews of the first class were misclassified as the second class (False Positives).

- 0 reviews of the first class were misclassified as the third class (False Positives).
- In the second row (for the Neutral class):
 - 0 reviews of the second class were misclassified as the first class (False Negatives).
 - 1,569 reviews of the second class were correctly classified as the second class (True Positives).
 - 0 reviews of the second class were misclassified as the third class (False Positives).

Figure 20

The GRU Confusion Measures



- In the third row (for the Positive class):
 - 63 instances of the third class were misclassified as the first class (False Negatives).
 - 488 instances of the third class were misclassified as the second class (False Negatives).

- 3,450 instances of the third class were correctly classified as the third class (True Positives).

Table 16 and Figure 21 illustrate the CNN-BI-LSTM performance examining COP28 for Arabic speech recognition.

Table 16

The CNN-BI-LSTM Performance

Dataset	Classes	Precision %	Recall %	F1-score %
COP28 Reviews	Negative	89	99	93
	Positive	100	91	94
	Neutral	83	100	88

The provided table outlines the performance metrics for the COP28 Reviews dataset across three distinct classes: Negative, Positive, and Neutral. These metrics, including precision, recall, and F1-score, offer valuable insights into the effectiveness of the classification model.

For the Negative class:

- The precision is 89%, indicating that 89% of the instances predicted as negative were indeed negative.

- The recall is 99%, implying that the model correctly identified 99% of the actual negative instances.

- The F1-score, which is the harmonic mean of precision and recall, is 93%, demonstrating a balanced performance in capturing both precision and recall for the Negative class.

For the Positive class:

- The precision is 100%, suggesting that all instances predicted as positive were indeed positive.

- The recall is 91%, signifying that 91% of the actual positive instances were correctly classified by the model.

- The F1-score is 94%, highlighting the overall effectiveness of the model in achieving a balance between precision and recall for the Positive class.

For the Neutral class:

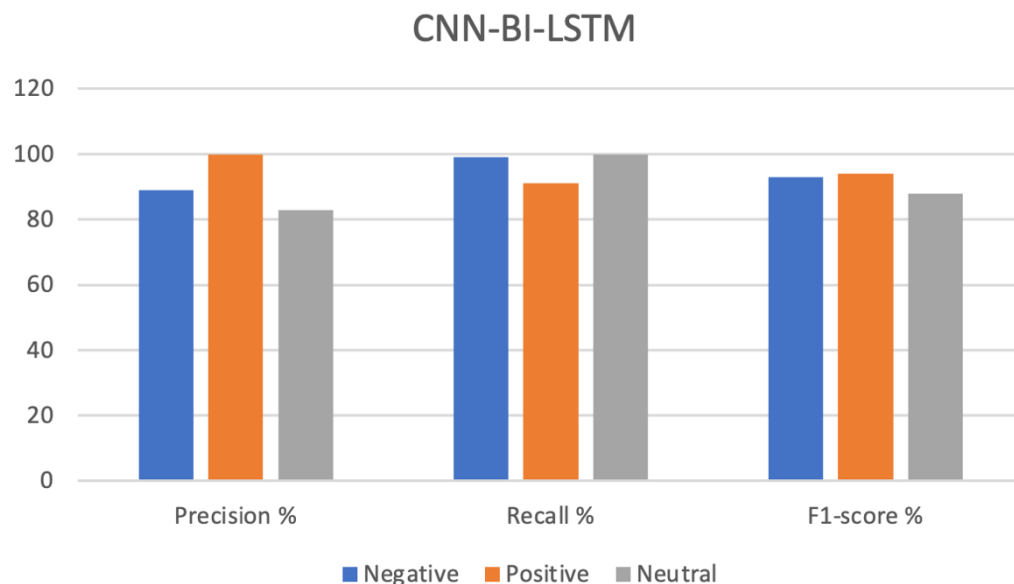
- The precision is 83%, indicating that 83% of the instances predicted as neutral were indeed neutral.

- The recall is 100%, demonstrating that the model correctly identified all actual neutral instances.

- The F1-score is 88%, showing a balanced performance in capturing both precision and recall for the Neutral class.

Figure 21

The CNN-BI-LSTM Performance



The provided confusion matrix as in Figure 22, offers a detailed representation of the classification performance of a model across multiple classes. Each row of the matrix corresponds to the actual class, while each column represents the predicted class.

Here is a breakdown of the confusion matrix:

- In the first row (for the Negative class):

- 172 reviews of the first class were correctly classified as the first class (True Negatives).

- 2 reviews of the first class were misclassified as the second class (False Positives).

- 0 reviews of the first class were misclassified as the third class (False Positives).

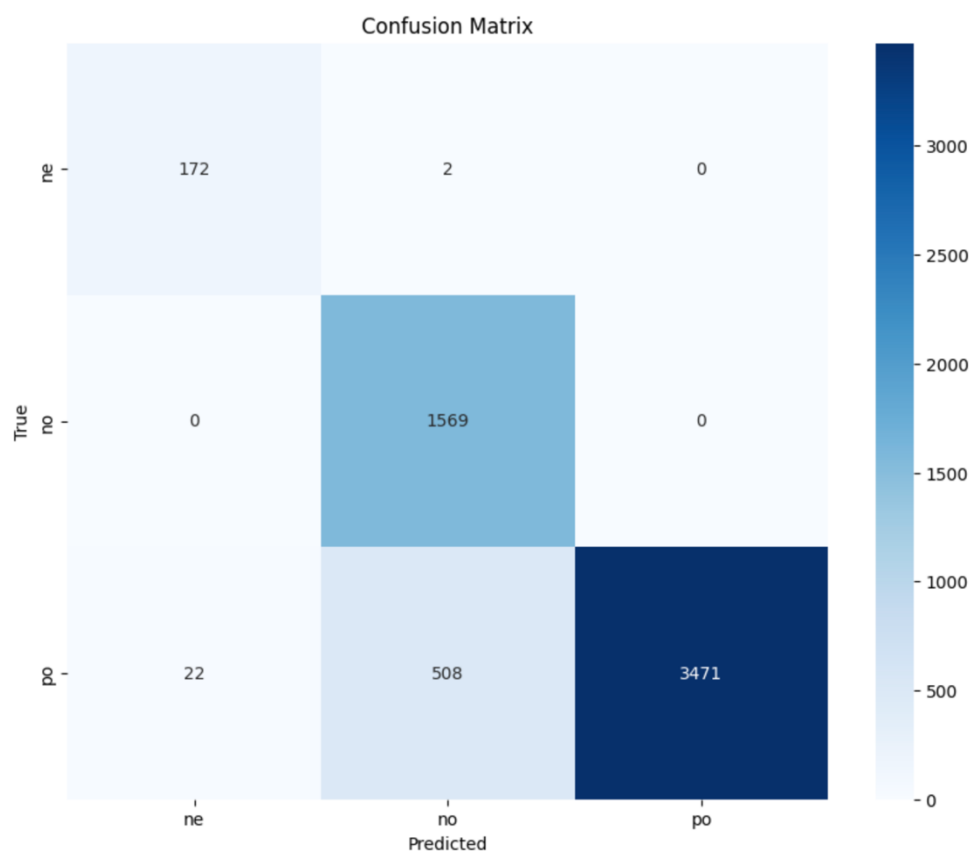
- In the second row (for the Neutral class):

- 0 reviews of the second class were misclassified as the first class (False Negatives).

- 1,569 reviews of the second class were correctly classified as the second class (True Positives).

Figure 22

The CNN-BI-LSTM Confusion Matrix



- 0 reviews of the second class were misclassified as the third class (False Positives).

- In the third row (for the Positive class):

- 22 instances of the third class were misclassified as the first class (False Negatives).

- 508 instances of the third class were misclassified as the second class (False Negatives).

- 3,471 instances of the third class were correctly classified as the third class (True Positives).

Table 17 and Figure 23 below propose the GRU-BI-LSTM hybrid model performance in Arabic speech recognition considering COP28 as a study case.

Table 17

The GRU-BI-LSTM Hybrid Model Performance

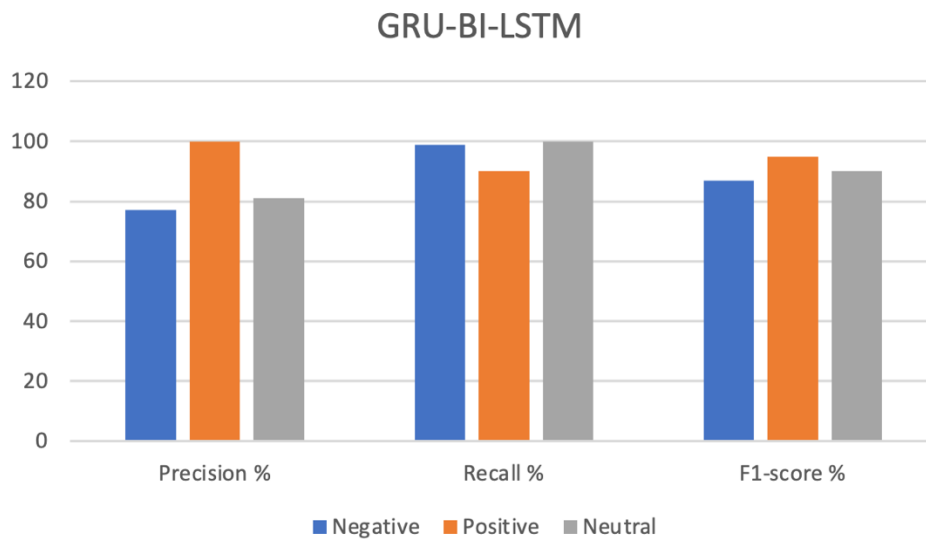
Dataset	Classes	Precision %	Recall %	F1-score %
COP28 Reviews	Negative	77	99	87
	Positive	100	90	95
	Neutral	81	100	90

The results presented in Table 18 and Figure 23 showcase the performance metrics of sentiment analysis for COP28 Reviews across three distinct classes: Negative, Positive, and Neutral.

For the Negative class:

- Precision is reported at 77%, indicating that out of all instances classified as negative, 77% were correctly classified.

- Recall, measuring the proportion of actual negative instances correctly identified by the model, is notably high at 99%.

Figure 23*The GRU-BI-LSTM Performance*

- The F1-score,

which balances precision and recall, is reported at 87%, reflecting the model's effectiveness in achieving a balance between correctly identifying negative instances and minimising false positives.

For the Positive class:

- Precision is at the highest level, standing at 100%, signifying that all instances classified as positive were indeed positive.

- Recall, indicating the proportion of actual positive instances correctly classified by the model, is reported at 90%.

- The F1-score, representing a harmonic mean of precision and recall, is reported at 95%, suggesting strong performance in capturing positive sentiments.

For the Neutral class:

- Precision is reported at 81%, indicating that 81% of instances classified as neutral were indeed neutral.

- Recall, denoting the proportion of actual neutral instances correctly classified by the model, is at 100%.

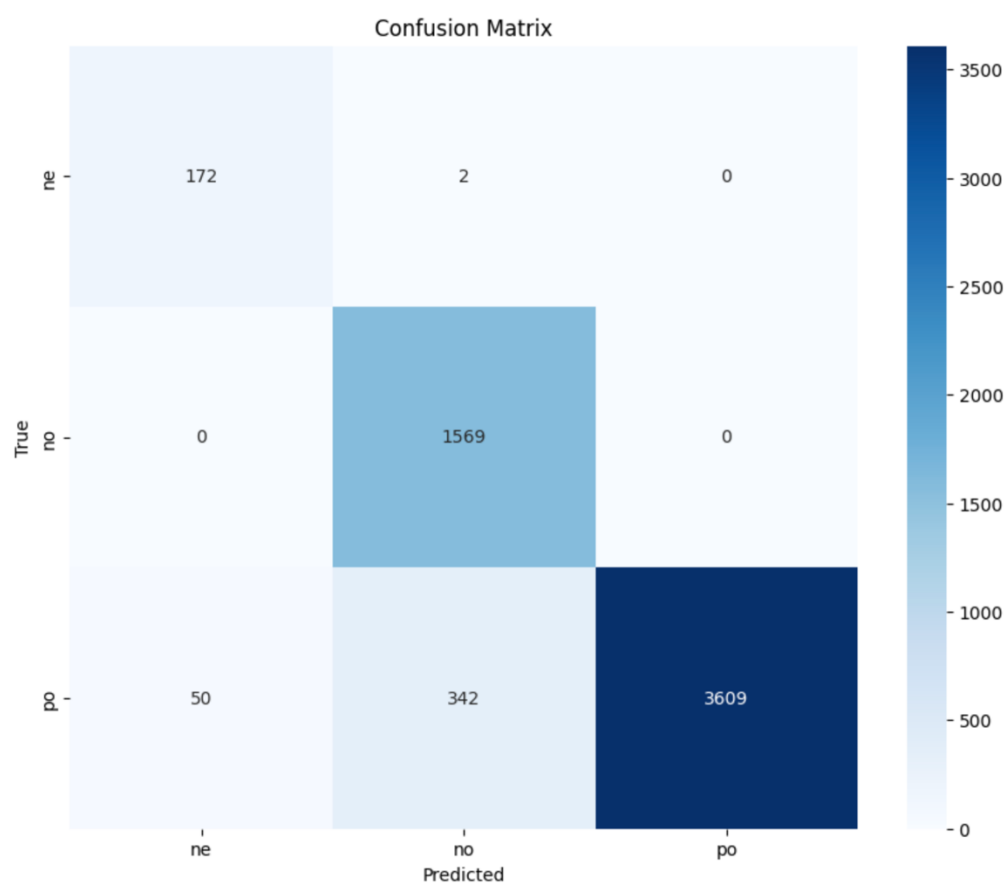
- The F1-score, reflecting the balance between precision and recall, is reported at 90%, demonstrating the model's effectiveness in accurately capturing neutral sentiments.

Overall, these results indicate a strong performance of the sentiment analysis model across all classes, with particularly notable performance in terms of precision and recall for the Positive and Neutral classes. The high recall rates across all classes suggest that the model effectively captures instances of each sentiment category, while the balanced F1-scores reflect the model's ability to maintain precision while maximising recall.

Figure 24 below shows the confusion matrix for the GRU-BI-LSTM. The provided confusion matrix offers a comprehensive overview of the classification performance of a sentiment analysis model across three distinct classes: Negative, Positive, and Neutral.

Figure 24

The Confusion Matrix of the GRU-BI-LSTM Hybrid Model



In the first row (Negative class):

- 172 instances of the Negative class were correctly classified as negative (True Negatives).

- 2 instances of the Negative class were misclassified as positive (False Positives).

- There were no instances of the Negative class being misclassified as neutral.

In the second row (Neutral reviews):

- All 1,569 instances of the Positive class were correctly classified as positive (True Positives).

- There were no instances of the Positive class being misclassified as negative or neutral.

In the third row (Positive reviews):

- 22 instances of the Neutral class were misclassified as negative (False Negatives).

- 508 instances of the Neutral class were misclassified as positive (False Negatives).

- 3,471 instances of the Neutral class were correctly classified as neutral (True Positives).

Overall, the confusion matrix provides insights into the model's performance, highlighting areas of correct and incorrect predictions across different classes. It serves as a valuable tool for evaluating the effectiveness of the sentiment analysis model and identifying areas for improvement.

These results provide a comprehensive understanding of the model's performance across different classes within the COP28 Reviews dataset, showcasing its ability to accurately classify instances into their respective categories. Table 18 and Figure 25 below show the accuracy of the classifiers in classifying the sentiment analysis of the Arabic speech recognition for the COP28 Reviews dataset.

Here is an overview of the results proposed in the table:

1. ****BI-LSTM****:

- Accuracy: 89%

- BI-LSTM is a type of RNN that processes input sequences in both forward and backward directions. Despite its effectiveness in capturing context information, the model achieves an accuracy of 89%.

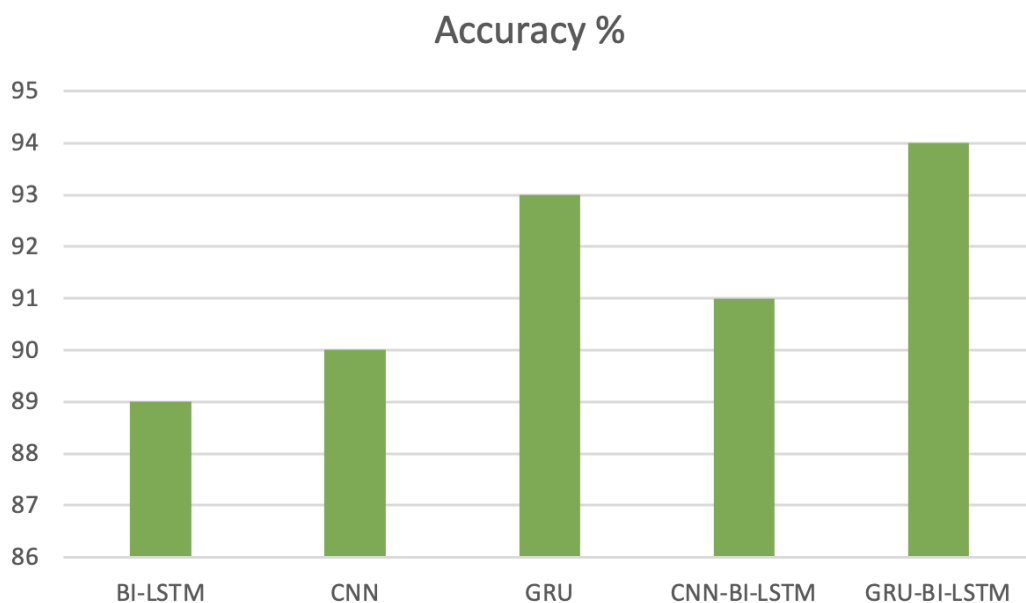
Table 18

Accuracy in the Classifiers

Model	Accuracy %
BI-LSTM	89
CNN	90
GRU	93
CNN-BI-LSTM	91
GRU-BI-LSTM	94

Figure 25

The Accuracy of the Classifiers in Arabic ASR



2. **CNN**:

- Accuracy: 90%

- CNNs are particularly effective in capturing spatial hierarchies in data, often applied to image data. In this context, it seems to perform slightly better than the basic BI-LSTM model, achieving an accuracy of 90%.

3. **GRU**:

- Accuracy: 93%

- GRU is a variant of RNN that aims to solve the vanishing gradient problem. It's known for its simpler architecture compared to LSTM. The GRU model achieves an accuracy of 93%, showing improved performance over the BI-LSTM and CNN models.

4. **CNN-BI-LSTM**:

- Accuracy: 91%

- This model combines the strengths of both CNN and BI-LSTM architectures. It likely uses CNN layers for feature extraction followed by BI-LSTM for sequence modelling. It achieves an accuracy of 91%, slightly higher than the standalone BI-LSTM model.

5. **GRU-BI-LSTM**:

- Accuracy: 94%

- This model combines the GRU and BI-LSTM architectures. GRU's ability to handle sequential data effectively combined with the context-capturing capabilities of BI-LSTM results in superior performance. It achieves the highest accuracy among all models listed, with 94%.

Overall, the GRU-BI-LSTM model outperforms the other models in terms of accuracy, indicating its effectiveness in capturing both sequential and contextual information in Arabic speech recognition, especially in the COP28 as a study case.

In the field of machine learning, choosing the structure for Arabic speech recognition, particularly in COP28 reviews, plays a vital role in achieving optimal results. Here is a

comparison of the network structures based on their accuracy metrics as outlined in Table 18 and Figure 25.

To begin with, we look at how the BI LSTM model performs, showing an accuracy rate of 89%. BI-LSTMs are skilled at capturing time-related patterns in data. However, in this case, it slightly lags behind models. Moving on to the CNN which achieves an accuracy of 90%, CNNs are strong at extracting patterns from data, making them particularly effective for tasks related to sound. Here, CNN's capability to capture characteristics gives it an edge over BI-LSTM. A significant improvement in accuracy is noted with the GRU model reaching an accuracy rate of 93%. GRUs, which are a version of LSTMs, display performance levels highlighting their effectiveness in handling sequential data. This substantial increase in accuracy underscores the importance of the GRU design. When directly comparing the BI LSTM and GRU models, it is clear that the GRU model performs better by a margin of 4%. This indicates that the GRU design, which combines forget and input gates, makes it more adaptable to the dataset or task at hand, resulting in improved performance. Similarly, when looking at the CNN and GRU models side by side, the GRU model outperforms CNN by 3% in terms of accuracy. This suggests that for this task or dataset, the GRU's ability to capture relationships is more beneficial than the CNN's strength in extracting spatial features.

In terms of combined architectures, the CNN-BI-LSTM model shows an improvement over using BI LSTM alone with an accuracy of 91%. This combination leverages both CNNs' and LSTMs' strengths, leading to a performance boost. Lastly, the GRU-BI-LSTM model stands out as the performer with an accuracy of 94%. By replacing LSTM layers with GRU layers in this combined architecture, it achieves the accuracy among all the models mentioned. This highlights how integrating GRU's simplicity with BI-LSTM's bidirectional modelling results in performance when capturing both sequential and contextual information. In summary, this detailed comparison sheds light on the variations in

performance between types of neural network structures. Each design has its advantages and drawbacks, and the decision on which model to use should be based on the given project's needs. The effectiveness of the GRU-BI-LSTM model emphasises how crucial architectural planning is in attaining highly effective results in machine learning assignments.

4.4 Conclusion and Future Work

Arabic, as one of the world's most commonly spoken languages, carries significant cultural, social, and economic importance via diverse regions. With over 310 million native speakers and an extensive diaspora, Arabic serves as a primary means of communication in numerous countries spanning the Middle East, North Africa, and beyond. In an era marked by rapid technological advancement, the development of Arabic speech recognition systems represents a pivotal frontier in enhancing accessibility, efficiency, and usability across various domains. Speech recognition, a branch of AI and NLP, enables machines to transcribe spoken language into text, facilitating seamless interaction between humans and technology. While considerable progress has been made in speech recognition for widely spoken languages such as English, challenges persist in effectively adapting these technologies to accommodate the unique characteristics of Arabic. These challenges stem from the language's rich morphology, diverse dialects, non-linear script, and contextual nuances, presenting formidable hurdles for conventional speech recognition systems. In this context, the investigation of Arabic speech recognition within the framework of COP28 serves as a compelling case study. COP28, an international gathering focused on addressing global environmental challenges and fostering sustainable development, attracts participants from diverse linguistic backgrounds, including Arabic speakers. The integration of speech recognition technologies at COP28 holds the potential to streamline communication, facilitate multilingual collaboration, and enhance accessibility for Arabic-speaking delegates and stakeholders. Through a comprehensive examination of various speech recognition models,

including CNN, BI-LSTM, GRU, and hybrid architectures such as CNN-BI-LSTM and GRU-BI-LSTM, valuable insights can be gleaned into the performance and efficacy of these models within the distinct context of Arabic speech recognition. By analysing key metrics such as precision, recall, and the F1-score across different sentiment categories, ranging from positive and negative to neutral, a nuanced understanding of each model's strengths and limitations can be attained. The BI-LSTM model achieves commendable performance across all classes, with precision, recall, and F1-score ranging from 73% to 100%. It exhibits particularly high recall rates for Negative and Neutral classes, indicating its ability to effectively capture instances of these sentiments. The CNN model showcases robust performance, especially in terms of precision and recall for the Negative and Positive classes, with scores above 78%. It demonstrates superior recall for all classes, indicating its proficiency in capturing relevant instances from the dataset. However, precision for the Neutral class is slightly lower, potentially due to the misclassification of other sentiments as neutral. The GRU model stands out with impressive precision and recall for the Negative class, achieving scores of 96% and 90%, respectively. It also performs well in the Positive class, with both precision and recall above 90%. However, its precision for the Neutral class is relatively lower compared to other models. The hybrid models, CNN-BI-LSTM and GRU-BI-LSTM, leverage the strengths of their constituent models to achieve competitive performance. Both hybrid models exhibit balanced precision and recall across classes, with F1 scores ranging from 87% to 95%. This highlights the potential of hybrid architectures to enhance the overall performance by integrating diverse features and representations. This dissertation endeavours to elucidate the significance of Arabic speech recognition within the broader landscape of

AI and NLP, contextualising its relevance through the lens of COP28 as a unique case study. By delving into the intricacies of model performance and exploring future directions

for improvement, this study aims to contribute to the advancement of Arabic speech recognition technology, ultimately fostering greater inclusivity, accessibility, and innovation in multilingual environments such as COP28 and beyond.

****Future Directions: ****

For future research endeavours, scholars may consider incorporating a variety of multimedia elements, involving images, videos, and textual data, into the investigation of COP28. By integrating these additional modalities, researchers can obtain a more comprehensive and nuanced understanding of the conference dynamics, discussions, and sentiments expressed by participants

Including images in the analysis allows researchers to capture visual cues, such as presentation slides, infographics, and photographs, which complement the spoken discourse and provide valuable contextual information. Visual content can offer insights into the topics discussed, the engagement level of the participants, and the overall atmosphere of the conference sessions. Analysing images alongside audio data enables researchers to correlate visual stimuli with spoken content, facilitating a deeper interpretation of conference proceedings.

Similarly, incorporating videos into the investigation enriches the dataset by capturing non-verbal communication cues such as facial expressions, gestures, and body language exhibited by speakers and attendees. Video recordings of conference sessions offer a holistic view of participant interactions, speaker dynamics, and audience reactions, which may not be fully captured through audio analysis alone. By analysing both audio and video data concurrently, researchers can gain a more nuanced understanding of communication patterns and sentiment dynamics within COP28.

Additionally, including textual data such as the transcripts of speeches, presentation slides, social media posts, and textual annotations enhances the depth and breadth of the

investigation. Textual content provides supplementary data on the topics discussed, key messages conveyed, and sentiments expressed by participants across different communication channels. Analysing textual data in conjunction with multimedia elements enables researchers to triangulate findings, validate interpretations, and uncover hidden insights within the conference discourse.

By embracing a multimodal approach to data collection and analysis, researchers can overcome the limitations of single-modality studies and achieve a more holistic perspective of COP28. Integrating images, videos, and texts into the investigation not only enriches the dataset but also fosters a deeper understanding of the complex interplay between verbal and non-verbal communication modalities within the conference context. This multimodal approach paves the way for future research endeavours aimed at unravelling the intricacies of COP28 and informing evidence-based decision-making processes in global environmental governance.

In conclusion, the investigation into COP28 for Arabic speech recognition underscores the importance of leveraging diverse models and methodologies to address the unique challenges and nuances of the Arabic language. By continuing to refine and innovate existing approaches, the field can pave the way for more accurate, reliable, and accessible Arabic speech recognition systems in the future.

4.5 Summary

We provided a deconstructed explanation of the research setup in this section of the dissertation, including the procedures and library resources that were used. In addition, a summary of the investigation's settings is suggested in this chapter, along with the initial processing, deep learning, and hybrid deep learning models' hyperparameter configurations.

References

- Abdel-Hamid L. Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features, *Speech Communication.*, vol. 122, pp. 19–30, Sep. 2020, doi: 10.1016/j.specom.2020.04.005
- Abdul-Mageed M., Diab M., Kübler S. SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media,” in *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, Korea: Association for Computational Linguistics, Jul. 2012, pp. 19–28. Accessed: Nov. 03, 2023. [Online]. Available: <https://aclanthology.org/W12-3705>
- Ahmad M., Aftab S., Muhammad S., Waheed U. Tools and Techniques for Lexicon Driven Sentiment Analysis: A Review . International journal of multidisciplinary sciences and engineering, vol. 8, no. 1, january 2017
- Ahmed A., Hifny Y., Shaalan K., Toral S.,. End-to-End Lexicon Free Arabic Speech Recognition Using Recurrent Neural Networks. pp. 231–248. 2018, https://doi.org/10.1142/9789813229396_0011
- Albawi S., Mohammed A., Al-Zawi S. Understanding of a convolutional neural network. Presented at the 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. <https://doi.org/10.1109/ICEngTechnol.2017.8308186>
- Al-Dujaili J., Ebrahimi-Moghadam. Speech Emotion Recognition: A Comprehensive Survey. *Wirel. Pers. Communication.* 129, 2525–2561. 2023, <https://doi.org/10.1007/s11277-023-10244-3>
- Ali A., Hifny Y., “Efficient Arabic Emotion Recognition Using Deep Neural Networks,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6710–6714. doi: 10.1109/ICASSP.2019.8683632.

- Alkhatib, M., Shaalan, K., 2017. Natural Language Processing for Arabic Metaphors: A Conceptual Approach, in: Hassanien, A.E., Shaalan, K., Gaber, T., Azar, A.T., Tolba, M.F. (Eds.), *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016, Advances in Intelligent Systems and Computing*. Springer International Publishing, Cham, pp. 170–181. https://doi.org/10.1007/978-3-319-48308-5_17
- Al-Malla A., Jafar A., Ghneim N. Pre-trained CNNs as Feature-Extraction Modules for Image Captioning: An Experimental Study. *ELCVIA Electron. Lett. Computer. Vis. Image Analysis* 2022. 21, 0001–0016. DOI:10.5565/rev/elcvia.1436
- Amin S., Uddin M., Al-Baity H., Zeb M., Khan M., “Machine Learning Approach for COVID-19 Detection on Twitter,” *Computer. Materials. Continuous.*, vol. 68, no. 2, pp. 2231–2247, 2021, doi: 10.32604/cmc.2021.016896.
- Angelidou M., “Smart cities: A conjuncture of four forces,” *Cities*, vol. 47, pp. 95–106, Sep. 2015, doi: 10.1016/j.cities.2015.05.004
- Arab Day for Population and Development October 28, 2023 | HPC [WWW Document],. URL <https://www.hpc.org.jo/en/content/arab-day-population-and-development-october-28-2023> (accessed 2.19.24).
- Arab world - total population 2011-2021 [WWW Document], . Statista. URL <https://www.statista.com/statistics/806106/total-population-arab-league/> (accessed 5.17.23).
- Baevski A., Zhou Y., Mohamed A., Auli M. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 12449–12460. Accessed: Feb. 19, 2024. [Online]. Available:

<https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html>

Bahdanau D., Cho K., Bengio Y. “Neural Machine Translation by Jointly Learning to Align and Translate,” May 19, 2016, *arXiv*: arXiv:1409.0473. doi: 10.48550/arXiv.1409.0473

Belinkov Y., Ali A., Glass J. “Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition,” Apr. 19, 2020, *arXiv*: arXiv:1907.04224. doi: 10.48550/arXiv.1907.04224

Bhavitha B., Rodrigues A., Chiplunkar N. Comparative study of machine learning techniques in sentimental analysis. 2017 International. Conference. Invention. Communication. Computer. Technology. ICICCT 216–221. 2017, <https://doi.org/10.1109/ICICCT.2017.7975191>

Cai, M., “Natural language processing for urban research: A systematic review,” *Heliyon*, vol. 7, no. 3, p. e06322, Mar. 2021, doi: 10.1016/j.heliyon.2021.e06322.

Carneiro T., Nóbrega R., Nepomuceno T., Bian G., Albuquerque V., Filho, P. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access* 6. 2018, <https://doi.org/10.1109/access.2018.2874767>

Casey A., Davidson E., Poon M., Dong H., Duma D., Grivas A., Grover C., Suárez-Paniagua V., Tobin R., Whiteley W., Wu H., Alex B. A systematic review of natural language processing applied to radiology reports. *BMC Medical Information. Decision Making*. 21, 179. 2021, <https://doi.org/10.1186/s12911-021-01533-7>

Cheng S. The Extended Technology Stack of Web 3.0, in: Cheng, S. (Ed.), *Web 3.0: Concept, Content and Context*. Springer Nature, Singapore, pp. 57–81. 2024 https://doi.org/10.1007/978-981-99-6319-5_3

Choubey V. Text classification using CNN. Voice Tech Podcast. 2020 URL

<https://medium.com/voice-tech-podcast/text-classification-using-cnn-9ade8155dfb9>

(accessed 1.5.23).

Chung J., Gulcehre C., Cho K., Bengio Y. Empirical Evaluation of Gated Recurrent Neural

Networks on Sequence Modeling. 2014, <https://doi.org/10.48550/arXiv.1412.3555>

COP28 - Carbon Market Watch [WWW Document]. URL

<https://carbonmarketwatch.org/cop28/>.

COP28 UAE - United Nations Climate Change Conference [WWW Document],. URL

<https://www.cop28.com/en/> (accessed 2.18.24).

Coppin G., Sander D. Chapter 1 - Theoretical approaches to emotion and its measurement, in:

Meiselman, H.L., Emotion Measurement (Second Edition). Woodhead Publishing, pp.

3–37. 2021, <https://doi.org/10.1016/B978-0-12-821124-3.00001-6>

Cummins N., Schuller B.W. Latest Advances in Computational Speech Analysis for Mobile

Sensing. Studies in Neuroscience, Psychology and Behavioral Economics 141–159.

2019 https://doi.org/10.1007/978-3-030-31620-4_9

Demographics of the Arab world, 2024. . Wikipedia.

https://en.wikipedia.org/wiki/Demographics_of_the_Arab_world

Dhouib A., Othman A., El Ghoul O., Khribi K., Al Sinani A., 2022. Arabic Automatic Speech

Recognition: A Systematic Literature Review. Applied Science. 12, 8898.

<https://doi.org/10.3390/app12178898>

Dong X., Lian Y. A review of social media-based public opinion analyses: Challenges and

recommendations. Technology Social. 67, 101724. 2021

<https://doi.org/10.1016/j.techsoc.2021.101724>

Dukes K., Habash N., “Morphological Annotation of Quranic Arabic,” in *Proceedings of the*

Seventh International Conference on Language Resources and Evaluation

- (*LREC'10*), European Language Resources Association (ELRA), Accessed: Aug. 12, 2024. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/276_Paper.pdf.
- Dutta Das D., Sharma S., Natani S., Khare N., Singh B. Sentimental Analysis for Airline Twitter data 263, 042067. 2017, <https://doi.org/10.1088/1757-899X/263/4/042067>
- Elnagar A., Al-Debsi R., Einea O. “Arabic text classification using deep learning models,” *Inf. Process. Management.*, vol. 57, no. 1, p. 102121, Jan. 2020, doi: 10.1016/j.ipm.2019.102121.
- Flemotomos N., Martinez V.R., Chen Z., Singla K., Ardulov V., Peri R., Caperton D., Gibson J., Tanana M.J., Georgiou P., Van Epps J., Lord S.P., Hirsch T., Imel Z.E., Atkins, D.C., Narayanan, S., 2022. Automated evaluation of psychotherapy skills using speech and language technologies. *Behave Res. Methods* 54, 690–711. <https://doi.org/10.3758/s13428-021-01623-4>
- Gao Y., Glowacka D. “Deep Gate Recurrent Neural Network,” in *Proceedings of The 8th Asian Conference on Machine Learning*, PMLR, Nov. 2016, pp. 350–365. Accessed: Nov. 12, 2023. [Online]. Available: <https://proceedings.mlr.press/v63/gao30.html>.
- Garouani M., Kharroubi J. MAC: An Open and Free Moroccan Arabic Corpus for Sentiment Analysis, *Innovations in Smart Cities Applications Volume 5*. 2022, Springer International Publishing, Cham, pp. 849–858. https://doi.org/10.1007/978-3-030-94191-8_68
- Gaye B., Wulamu A. Sentiment Analysis of Text Classification Algorithms Using Confusion Matrix, in: Ning, H. (Ed.), *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health, Communications in Computer and Information Science*. Springer Singapore, Singapore, 2019, pp. 231–241. https://doi.org/10.1007/978-981-15-1922-2_16

Gu J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., Liu T., Wang X., Wang G., Ca J.,

Chen T. Recent advances in convolutional neural networks. 2018 Pattern Recognit.

77, 354–377. <https://doi.org/10.1016/j.patcog.2017.10.013>

Guetterman T.C., Chang T., DeJonckheere M., Basu T., Scruggs E., Vydiswaran V.

Augmenting Qualitative Text Analysis with Natural Language Processing:

Methodological Study. J. Med. 2018 Internet Res. 20, e231.

<https://doi.org/10.2196/jmir.9702>

Hema C., Garcia Marquez. Emotional speech Recognition using CNN and Deep learning

techniques. Appl. Acoustic. 2023, 211, 109492.

<https://doi.org/10.1016/j.apacoust.2023.109492>

Hochreiter S., Schmidhuber J. Long Short-Term Memory. Neural Computer. (2016) 9, 1735–

1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Irie K., Tüske Z., Alkhouli T., Schlüter R., Ney H. LSTM, GRU, Highway and a Bit of

Attention: An Empirical Overview for Language Modeling in Speech Recognition.

Processing. Interspeech 2016, 3519-3523, doi: 10.21437/Interspeech.2016-491Joyce

B., Deng J. Sentiment analysis of tweets for the 2016 US presidential election, in:

2017 IEEE MIT Undergraduate Research Technology Conference (URTC). Presented

at the 2017 IEEE MIT Undergraduate Research Technology Conference (URTC),

IEEE, Cambridge, MA, pp. 1–4. <https://doi.org/10.1109/URTC.2017.8284176>

Kanakaraj M., Guddeti R. Performance analysis of Ensemble methods on Twitter sentiment

analysis using NLP techniques, in: Proceedings of the 2015 IEEE 9th International

Conference on Semantic Computing (IEEE ICSC 2015). Presented at the Proceedings

of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC

2015), pp. 169–170. <https://doi.org/10.1109/ICOSC.2015.7050801>

- Karami A., Bennett L.S., He X. Mining Public Opinion about Economic Issues: Twitter and the U.S. Presidential Election. *Int. J. Strategic Decision Science* 2018. 9, 18–28.
<https://doi.org/10.4018/IJSDS.2018010102>
- Keshavarz H., Abadeh M.S. Accurate frequency-based lexicon generation for opinion mining. *J. Intelligent* 2017. *Fuzzy Syst.* 33, 2223–2234. <https://doi.org/10.3233/JIFS-16562>
- Khairi N., Mohamed A., Yusof N. Feature Selection Methods in Sentiment Analysis: A Review. pp. 1–7. 2020 <https://doi.org/10.1145/3386723.3387840>
- Latif S., Qadir J., Qayyum A., Usama M., Younis S. Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. *IEEE Rev. Biomed. Eng.* 14, 342–356. <https://doi.org/10.1109/RBME.2021.3006860>
- Ligthart A., Catal C., Tekinerdogan B. Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligent.* 2021 *Rev.* 54, 4997–5053. <https://doi.org/10.1007/s10462-021-09973-3>
- Maurya C.G., Jha S.K. Sentiment Analysis: A Hybrid Approach on Twitter Data. *Procedia Computer. Science., International Conference on Machine Learning and Data Engineering (ICMLDE 2024)* 235, 990–999.
<https://doi.org/10.1016/j.procs.2024.04.094>
- Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering. J.* 5, 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Meftah A., Qamhan M., Alotaibi Y.A., Zakariah M.,. Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus. *International. J. Simulation. Systems. Science. Technology* 2020. <https://doi.org/10.5013/IJSSST.a.21.02.21>
- Mishra S., Bhatnagar N. Speech emotion recognition and classification using hybrid deep CNN and Bi-LSTM model 2023. *Multimedia. Tools Appl.*
<https://doi.org/10.1007/s11042-023-16849-x>

- Mourad A., Darwish K. Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs, in: Balahur, A., van der Goot, E., Montoyo, A. (Eds.), Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Presented at the WASSA 2013, Association for Computational Linguistics, Atlanta, Georgia, pp. 55–64.
- Mozetič, I., Torgo, L., Cerqueira, V., Smailović, J., 2018. How to evaluate sentiment classifiers for Twitter time-ordered data? PLoS ONE 13, e0194317.
<https://doi.org/10.1371/journal.pone.0194317>
- Munir, H.S., Ren, S., Mustafa, M., Siddique, C.N., Qayyum, S., 2021. Attention based GRU-LSTM for software defect prediction. PLOS ONE 16, e0247444.
<https://doi.org/10.1371/journal.pone.0247444>
- Naidu, G., Zuva, T., Sibanda, E.M., 2023. A Review of Evaluation Metrics in Machine Learning Algorithms, in: Silhavy, R., Silhavy, P. (Eds.), Artificial Intelligence Application in Networks and Systems. Springer International Publishing, Cham, pp. 15–25. https://doi.org/10.1007/978-3-031-35314-7_2
- Olivier, R., Raj, B., 2023. There is more than one kind of robustness: Fooling Whisper with adversarial examples. <https://doi.org/10.48550/arXiv.2210.17316>
- O'Shea, K., Nash, R., 2015. An Introduction to Convolutional Neural Networks.
<https://doi.org/10.48550/arXiv.1511.08458>
- Ottom, M.A., Nahar, K.M.O., 2021. Social Media Sentiment Analysis: The Hajj Tweets Case Study. J. Comput. Sci. 17, 265–274. <https://doi.org/10.3844/jcssp.2021.265.274>
- Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2015 IEEE International

- Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210.
<https://doi.org/10.1109/ICASSP.2015.7178964>
- Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Found. Trends® Inf. Retr.* 2, 1–135. <https://doi.org/10.1561/15000000011>
- Prastyo, P.H., Ardiyanto, I., Hidayat, R., 2020. A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods, in: 2020 6th International Conference on Science and Technology (ICST). Presented at the 2020 6th International Conference on Science and Technology (ICST), pp. 1–6.
<https://doi.org/10.1109/ICST50505.2020.9732885>
- Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust Speech Recognition via Large-Scale Weak Supervision.
- Rajeswari, A.M., Mahalakshmi, M., Nithyashree, R., Nalini, G., 2020. Sentiment Analysis for Predicting Customer Reviews using a Hybrid Approach. 2020 *Adv. Computer. Communication. Technology. High Perform. Appl. ACCTHPA* 200–205.
<https://doi.org/10.1109/ACCTHPA49271.2020.9213236>
- Ramanathan V., Meyyappan T., 2019. Twitter Text Mining for Sentiment Analysis on People’s Feedback about Oman Tourism. 2019 4th MEC Int. Conf. Big Data Smart City ICBDS C 1–5. <https://doi.org/10.1109/ICBDSC.2019.8645596>
- Ray S.K., Shaalan K., 2016. A Review and Future Perspectives of Arabic Question Answering Systems. *IEEE Trans. Knowl. Data Eng.* 28, 3169–3190.
<https://doi.org/10.1109/TKDE.2016.2607201>
- Rousseau A., Deléglise P., Estève Y., 2012. “TED-LIUM: an Automatic Speech Recognition dedicated corpus,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 125–129. Accessed: Feb.

- 19, 2024. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/698_Paper.pdf.
- Saxena U., Mishra S., Rodrigo H., Choudhury M., 2022. Functional consequences of extended high frequency hearing impairment: Evidence from the speech, spatial, and qualities of hearing scale. *J. Acoust. Soc. Am.* 152, 2946–2952.
<https://doi.org/10.1121/10.0015200>
- Scheurer J., Campos J.A., Korbak T., Chan J.S., Chen A., Cho K., Perez E., 2024. Training Language Models with Language Feedback at Scale.
<https://doi.org/10.48550/arXiv.2303.16755>
- Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. <https://doi.org/10.1109/78.650093>
- Shaalán K., Siddiqui S., Alkhatib M., Abdel Monem A., 2018. Challenges in Arabic Natural Language Processing, in: *Computational Linguistics, Speech and Image Processing for Arabic Language*. World Scientific, pp. 59–83.
https://doi.org/10.1142/9789813229396_0003
- Shoufan, A., Alameri, S., 2015. Natural language processing for dialectal Arabic: A survey, in: *Proceedings of the Second Workshop on Arabic Natural Language Processing*. pp. 36–48.
- Tartir, S., Abdul-Nabi, I., 2017. Semantic Sentiment Analysis in Arabic Social Media. *J. King Saud Univ. - Comput. Inf. Sci.*, *Arabic Natural Language Processing: Models, Systems and Applications* 29, 229–233. <https://doi.org/10.1016/j.jksuci.2016.11.011>
- Vashisht, S., Kumar, P., Trivedi, M.C., 2024. Enhanced GRU-BiLSTM Technique for Crop Yield Prediction. *Multimed. Tools Appl.* <https://doi.org/10.1007/s11042-024-18898-2>
- Wang, D., Wang, X., Lv, S., 2019. An Overview of End-to-End Automatic Speech Recognition. *Symmetry* 11, 1018. <https://doi.org/10.3390/sym11081018>

Ware, J., 2020. Testament to Murder: The Violent Far-Right's Increasing Use of Terrorist Manifestos. International Centre for Counter-Terrorism.

We need transformative policies and actions to stand a chance of winning the race to net zero.

[WWW Document], 2023. . World Econ. Forum. URL

<https://www.weforum.org/agenda/2023/10/alliance-of-ceo-climate-leaders-open-letter-cop28/> (accessed 5.7.24).

Weger, R., Lossio-Ventura, J.A., Rose-McCandlish, M., Shaw, J.S., Sinclair, S., Pereira, F., Chung, J.Y., Atlas, L.Y., 2023. Trends in Language Use During the COVID-19 Pandemic and Relationship Between Language Use and Mental Health: Text Analysis Based on Free Responses From a Longitudinal Study. *JMIR Ment. Health* 10, e40899. <https://doi.org/10.2196/40899>

Xiao, J., Han, L., Zhang, H., 2022. Exploring Driving Factors of Digital Transformation among Local Governments: Foundations for Smart City Construction in China. *Sustainability* 14, 14980. <https://doi.org/10.3390/su142214980>

Zerari, N., Abdelhamid, S., Bouzgou, H., Raymond, C., 2019. Bidirectional deep architecture for Arabic speech recognition. *Open Comput. Sci.* 9, 92–102. <https://doi.org/10.1515/comp-2019-0004>

Zhang, S., Zheng, D., Hu, X., Yang, M., 2015. Bidirectional Long Short-Term Memory Networks for Relation Classification, in: Zhao, H. (Ed.), *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. Presented at the PACLIC 2015, Shanghai, China, pp. 73–78.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., Meng, Z., Hu, K., Rosenberg, A., Prabhavalkar, R., Park, D.S., Haghani, P., Riesa, J., Perng, G., Soltan, H., Strohman, T., Ramabhadran, B., Sainath, T., Moreno, P., Chiu, C.-C., Schalkwyk, J., Beaufays, F., Wu, Y., 2023. Google USM: Scaling

Automatic Speech Recognition Beyond 100 Languages.

<https://doi.org/10.48550/arXiv.2303.01037>