

A Computational Linguistic Approach to Gender Analysis and Classification of Kuwaiti Arabic



Hesah Aldihan

Department of Computer Science
University of Sheffield

This dissertation is submitted for the degree of
Doctor of Philosophy

September 2024

I dedicate this thesis to my beloved grandmothers, Noura and Safia. May their souls rest in peace. I hold your memories dear and wish you were beside me, celebrating this achievement ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Hesah Aldihan
September 2024

Acknowledgements

As the African proverb goes: “If you want to go fast, go alone. If you want to go far, go together”. My doctoral journey, with all its challenges and triumphs has been a testament to the truth of these words. I have grown to deeply understand and appreciate the blessing of unwavering support and guidance, without which this milestone would not have been reached.

My sincere gratitude goes to my first supervisor, Prof. Rob Gaizauskas. I have truly been blessed to be guided and mentored by Rob throughout this journey. I am grateful to the wealth of knowledge I have gained from him, as he never hesitated to share valuable insights and suggestions to enhance my work. Rob to me is an inspiration in his dedication to research excellence. His commitment to guiding and supporting others has inspired me to pay it forward by emulating his example. I am also indebted to my second supervisor, Prof. Susan Fitzmaurice, the Dean of Humanities and Arts. Despite her demanding position and her hectic schedule, she consistently made time for me, providing guidance and assistance whenever needed. Her insights have greatly enriched this thesis and I feel fortunate to have had her support. I will always cherish Susan’s warm smile during our meetings and her kind, welcoming demeanor. Her positive energy and genuine kindness have left a lasting impression on me, and I am truly grateful for her presence throughout our collaboration. I will forever be grateful to Rob and Susan. Their combined expertise and perspectives from different fields have contributed immensely to the development of this thesis.

To my beloved father, Dr. Salah Aldihan. Before I embarked on this journey, the mere thought of being apart for four years troubled you deeply. However, life had its own plans, and it led me to spend six years abroad. Throughout this time, you remained the unwavering beacon of light in my life. Your phone calls after each meeting, your words of encouragement during moments of doubt, and your unlimited belief in my abilities have been the driving force behind my perseverance. I am eternally grateful for your sacrifices, your guidance, and your endless love.

To my loving mother, Rehab Alsubaiei. This journey has not been easy. There were moments when I struggled to find my footing, yet you were always there. No words can do

you justice for all the help you provided, all the nights you spent by my side, and the quiet, comforting whispers of your prayers that soothed my heart. I love you beyond measures now and always.

To my second mothers Awatif and Eqbal, my respected uncles Ahmed, Muneer, Khalid, and Salem, my beloved brothers Abdulrazaq and Muhammed, and the beautiful addition to our family my sister-in-law Muneera, thank you all for your support and belief in me. Every moment spent with you is treasured. Thank you for filling my days with warmth and happiness.

To Azoozi and Saloohi, my adored nephews, though distance may have kept us apart during some of the most precious moments of your childhood, your presence in my life has brought boundless joy and love, enriching my heart in ways I never imagined. knowing that I have two adorable nephews cheering me on makes every step forward feel all the more worthwhile. I look forward to the days ahead, filled with the promise of sharing life's adventures and weaving tales of my journey with you both. You bring light to my life, and I cherish every moment we spend together, making memories that will last a lifetime.

I am especially grateful for the inspiring minds I met at the Computer Science Department at Sheffield University who supported me throughout this journey. Special thanks to my colleagues and friends Hanan Halawani and Erfan Loweimi for their support from the beginning until the end. Your encouragement and belief in me have been truly invaluable.

A special acknowledgment to my colleague and best friend, Tarfah Alrashid, who became like a sister to me. Tarfah, you were the best thing that happened during this journey. I am incredibly grateful for your presence and for the lifetime best friend I gained in you.

I am also deeply thankful to my friends who made my time at Sheffield University enjoyable and memorable. Many thanks to Omaira, Ashwaq, Abla, Maha, Nora, Shurooq, Areej, and Amal. Thank you for the beautiful memories in Sheffield. Thanks to my colleagues at the NLP department Varvara, Sara and Reem.

Finally, and of utmost importance, I extend my sincere gratitude to Kuwait University for their sponsorship and for selecting me to pursue my PhD in Computational Linguistics. I am deeply thankful for this opportunity, which has been instrumental in shaping my academic journey. I also want to extend my gratitude to Dr. Mohammed Bin Naser for his belief in me and for encouraging me to pursue this degree and my heartfelt appreciation to my colleagues at Kuwait University for their invaluable contributions to the experimental work and their support throughout this journey. I'm deeply grateful to Hala Al-Sammar, Noor Algharabali, and Dhuha Al-Askar.

Abstract

This thesis focuses on the collection and computational analysis of Kuwaiti Arabic to test different sociolinguistic hypotheses related to gendered language use in social media. As the Kuwaiti Arabic dialect has some unique linguistic features that are stereotypically associated with gendered language usage, the study adopts a computational approach to study these features and draw insights into the relationship between language and gender in Kuwaiti Arabic. It contributes to the field of Arabic Natural Language Processing by providing three publicly available datasets: the Kuwaiti Arabic Gender-labelled WhatsApp Dataset (KAGen), the Kuwaiti Arabic Conversational Function WhatsApp Dataset (KACD), and the Kuwaiti Arabic Twitter Dataset (KATD).

The thesis unfolds in three main studies. The first study introduces the collection of KAGen, a Kuwaiti Arabic dataset that consists of WhatsApp exchanges of mixed gender Kuwaiti users collected from WhatsApp reading club groups that moved online during COVID-19. This dataset has been used to analyse different interactional and linguistic features to get insights about gender indicative features to inform the development of a gender classification system for Kuwaiti Arabic. The features studied are analysed quantitatively and qualitatively and are tested in a basic gender classification system trained and tested on the dataset.

The second study involves the development of an annotation framework to annotate the KAGen dataset according to the conversational functions employed in the turns. The study adopts an inductive thematic analysis approach to scrutinise the dataset and create the conversational function taglist that is used by annotators along with annotation guidelines to annotate the dataset which we name KACD. To the best of our knowledge, KACD is the only publicly available dataset of conversational Kuwaiti Arabic tagged according to conversational functions. The study provides insights regarding conversational function patterns and presents statistics on the distribution of conversational functions among men and women. Additionally, it offers qualitative observations, shedding light on distinctive linguistic patterns observed in the language used.

The third study aims to use machine learning models to automatically predict the gender of Kuwaiti Arabic social media users. For this study, KATD, a large publicly available dataset of Kuwaiti Arabic tweets is collected and labelled according to the gender of users. Two supervised learning approaches are taken to build the gender classification systems: a feature engineering approach and a deep learning approach. The feature engineering approach is adopted to test different linguistic features including the features inferred from the previous two studies to analyse their performance in predicting the gender of users. The deep learning approach, using pre-trained Transformer models, is also tested to assess how well pre-trained large language models perform in predicting the gender of Kuwaiti Arabic social media users.

Contents

List of Figures	15
List of Tables	17
1 Introduction	1
1.1 Overview of Thesis Contributions	2
1.2 Overview of the Thesis	4
2 Background	7
2.1 Computational Sociolinguistics	7
2.2 Language and Gender	8
2.2.1 Interdisciplinary Perspectives on Gender and Language	9
2.3 Language and Gender Studies on the Arabic Language	10
2.4 Kuwait: Demographics, Globalisation, and the Internet	11
2.5 Language in Kuwait	13
2.6 Language, Gender, and Social Media in Kuwait	13
2.7 Arabic	15
2.7.1 Classical Written Arabic	16
2.7.2 Modern Standard Arabic	16
2.7.3 Dialectal Arabic	16
2.8 Arabic Datasets for Computational Processing	16
2.8.1 Classical Arabic Datasets	17
2.8.2 Modern Standard Arabic Datasets	17
2.8.3 Dialectal Arabic Datasets	18
2.8.4 Kuwaiti Arabic Datasets for Natural Language Processing	19
2.9 Challenges in Preprocessing the Arabic Language	20
2.9.1 The Arabic Script	20
2.9.2 Rich Morphology of the Arabic Language	20

2.10	Arabic Preprocessing Tools	23
2.11	Comparative Analysis of Existing Arabic Preprocessing Tools	24
2.11.1	Tokenisation, Removing Diacritics, and Stop-word Removal	25
2.11.2	POS Tagging	26
2.11.3	Summary or Analysis	28
2.12	Text Classification	28
2.12.1	Statistical Approaches with Feature Engineering	29
2.12.2	Deep Learning Approaches with Transformers	36
2.13	Summary	40
3	Research Methods and Data	43
3.1	Research Questions	43
3.2	Research Methodology	44
3.3	Collection and Analysis of features from a WhatsApp Dataset	44
3.4	WhatsApp Dataset Annotation for Conversational Analysis	45
3.5	Gender Classification Using KA Tweets	46
3.5.1	Feature Engineering Approach	46
3.5.2	Deep Learning Approach	47
4	Collection and Computational Analysis of Linguistic Differences Amongst Men and Women in a Kuwaiti Arabic WhatsApp Dataset	49
4.1	Introduction	49
4.2	Related Work	50
4.3	Methodology	52
4.3.1	Data Collection	52
4.3.2	Data Preprocessing	53
4.3.3	Feature Analysis	54
4.4	Quantitative and Qualitative Analysis of Chats	55
4.4.1	Quantitative Analysis	55
4.4.2	Qualitative Analysis	62
4.5	A Baseline Gender Classification System	66
4.5.1	Data Preprocessing	67
4.5.2	Feature Extraction	67
4.5.3	Results and Analysis	69
4.6	Summary	70

5	WhatsApp Data Annotation for Conversational Analysis	71
5.1	Introduction	71
5.2	Related Work	72
5.3	WhatsApp Data Annotation Methodology	74
5.3.1	Context of the WhatsApp Reading Club Groups	74
5.3.2	The Development of the Conversational Function Tags and Guide- lines for Conversation Annotation	74
5.3.3	Recruitment and Training of Annotators	76
5.3.4	The Kuwaiti Arabic Conversational Function Dataset (KACD)	77
5.4	Results and Analysis	78
5.4.1	The Inter-Annotator Agreement Scores	79
5.4.2	Descriptive Statistics	80
5.4.3	Quantitative Analysis	80
5.5	Summary	87
6	Gender Classification Using Kuwaiti Arabic Tweets:	89
6.1	Introduction	89
6.2	Related Work	90
6.3	Data Collection	91
6.4	In Depth Analysis of Gender-indicative Vocabulary and Emojis	93
6.5	Feature Engineering Approach to Gender Classification	99
6.5.1	Classifier Development	100
6.5.2	Tweet Pre-processing	100
6.5.3	Feature Exploration	101
6.5.4	Feature Selection Methods	117
6.5.5	Combined Feature Experiments	121
6.6	Deep Learning Approach to Gender Classification	123
6.6.1	Models and Hyper-parameters	124
6.6.2	Accuracy Results Using Transformer Models	125
6.7	Analysis and Discussion	125
6.7.1	Performance Results Analysis	125
6.7.2	Failure Analysis	128
6.8	Summary	130
7	Conclusion and Future Work	133
7.1	Summary of Contributions	133
7.1.1	Kuwaiti Arabic Datasets	133

7.1.2	Gender Classification Systems	134
7.1.3	Insights into Gendered Linguistic Interactions in KA	134
7.2	Future Work	137
7.3	Concluding Remarks	138
Bibliography		139

List of Figures

2.1	The Arabic Alphabet with and without Diacritics. Reproduced from https://www.shamsaat.com/printables_packages	21
2.2	An Example of One Arabic Word and its' English Equivalent Represented by the Same Color	22
2.3	Comparison between the Output of Three Arabic Tokenisers	26
2.4	Text Classification Process. Figure from Bird et al. (2009)	30
2.5	SVM mapping the data points from the input space to a higher dimensional feature space to linearly separate the data. Figure from Luts et al. (2010).	31
2.6	The left figure illustrates various hyperplanes capable of separating the data, while the right figure demonstrates how SVMs construct a hyperplane that maximizes the margin between the different classes of data. Figure from Luts et al. (2010).	31
2.7	KNN Classifier from Imandoust et al. (2013)	34
2.8	Setting K to Different Values. Figure from https://www.analyticsvidhya.com/blog/2018/03/introduction-to-k-neighbours-algorithm-clustering/	35
2.9	Structure of a Decision Tree. Figure from Charbuty and Abdulazeez (2021)	36
2.10	Information Flow in a Self-attention Model. Figure from Jurafsky and Martin (2023).	37
2.11	Bidirectional information flow in a self-attention model. Figure from Jurafsky and Martin (2023).	39
2.12	Sequence Classification Using BERT. Figure from Jurafsky and Martin (2023).	40
4.1	Comparison of total number of turns before and after treating outliers	57
4.2	Distribution of Square Root Transformation Applied to Number of Turns	57
4.3	Distribution of Square Root Transformation Applied to Word Counts Divided by Number of Turns	59

4.4	Distribution of Square Root Transformation Applied to Emoji Counts Divided by Number of Turns	61
4.5	Most Frequent Words Used by Women	66
4.6	Most Frequent Words Used by Men	66
6.1	Distribution of Tweet Lengths for Females and Males	93
6.2	Characteristic Emojis of Females and Males Kuwaitis Twitter Users	98
6.3	Classifier Training	100
6.4	Classifier Testing	100
6.5	Example Input Sequence to a Transformer Based Model.	123
6.6	Confusion Matrix for Top 7 Features and TFIDF on (UP) Tweets	128
6.7	Confusion Matrix for Characteristic Vocabulary on (UP) Tweets	130

List of Tables

2.1	Comparison between the Output of Four Different Arabic POS Taggers for the sentence that translates to: “Corona Virus is spreading in more than 200 countries around the world, after it started spreading in China.”	27
4.1	Descriptive Statistics of the KAGen Dataset	53
4.2	Descriptive Statistics of the Features Analysed	55
4.3	Test of Normality for Square Root Transformed Turn Counts	58
4.4	T-Test Results for Square Root Transformed Turn Counts	58
4.5	Test of Normality for Square Root of Word Counts Divided by Number of Turns	59
4.6	Mann-Whitney Statistics for Square Root of Word Counts Divided by Turns	60
4.7	Test of Normality for Square Root of Emoji Counts Divided by Number of Turns	61
4.8	T-Test Results for Square Root Transformed Emoji Counts	62
4.9	Top Ten Emojis Used by Kuwaiti Women and Men	63
4.10	Accuracy (Acc) and Balanced Accuracy (BAcc) Results for the Baseline Gender Classification System Using 10 Fold Cross-validation. PP represents pre-processed text and UP represents unprocessed text. Comb.1 = sent len + stretched words + emoji bigrams. Comb.2 = word count + sent len + stretched words + emoji bigrams.	69
5.1	Conversational Tags and their Description	76
5.2	Descriptive Statistics for the KACD Corpus, Including Sentence and Word Counts per Tag and Gender	78
5.3	Sentence Level Inter-annotator Agreement Scores	79
5.4	Descriptive Statistics for the 7 Conversational Functions	80
5.5	Normality Test Results for the Normalised Proportions of Tag Usage	81

5.6	Hypotheses and Statistical Tests Used to Compare Men and Women Across their Usage of Conversational Functions	82
5.7	T-Test Results for Arranging Club Meeting	82
5.8	Mann-Whitney U Test Results for Arranging Club Meeting	83
5.9	Mann-Whitney U Test Results for Book Discussion	84
5.10	Mann-Whitney U Test Results for Feedback in Club Meeting	84
5.11	Mann-Whitney U Test Results for General Reading-related Discussion . . .	85
5.12	Mann-Whitney U Test Results for Greeting	85
5.13	T-Test Results for Social Interaction	86
5.14	Mann-Whitney U Test Results for Social Interaction	86
6.1	Descriptive Statistics of the Kuwaiti Arabic Gender-labelled Dataset (KATD)	92
6.2	Contingency table of the observed frequencies.	94
6.3	Words Characteristic of Kuwaiti Female Language	97
6.4	Words Characteristic of Kuwaiti Male Language	97
6.5	Accuracy Results on Pre-processed (PP) and Unprocessed (UP) Word Count Feature	103
6.6	Accuracy Results on Pre-processed (PP) and Unprocessed (UP) TFIDF Feature	104
6.7	Accuracy results of classifiers using the StretchedWordsBinary and StretchedWordsCount feature sets on the unprocessed dataset.	105
6.8	Accuracy Results on Pre-processed (PP) and Unprocessed (UP)Feature Sets	106
6.9	Results of Classifiers on the Punctuation Count Feature Using the Unprocessed Dataset.	107
6.10	Accuracy Results of Classifiers on Different Variations of Exclamation Mark Features	108
6.11	Accuracy Results on Pre-processed (PP) and Unprocessed (UP) POS Counts	108
6.12	Accuracy Results on Pre-processed (PP) and Unprocessed (UP) Total Adjective Counts Feature	109
6.13	Accuracy Results of Classifiers on Different Variations of Code-switching Features	110
6.14	Accuracy Results on Pre-processed (PP) and Unprocessed (UP) Word Embeddings	111
6.15	Accuracy Results of Classifiers on Different Variations of Characteristic Vocabulary Features (Pre-processed (PP) and Unprocessed (UP))	112
6.16	Accuracy Results on Pre-processed (PP) and Unprocessed (UP) Sentiment Analysis Scores.	114
6.17	Classifier Accuracy for Different Emojis Features	116

6.18	Accuracy Results of Classifiers on URL Usage	116
6.19	Summary of Features Performance Results.	118
6.20	Accuracy Results of Different Feature combinations using Mutual Information on Unprocessed (UP) and Preprocessed (PP) Tweets	120
6.21	Comparison of Accuracy Scores for Features with P-Values Less then 0.05 According to ANOVA F-test Feature Selection Method Using Unprocessed (UP) and Preprocessed (PP) Tweets	121
6.22	Comparison of Accuracy Scores for Different Feature Combinations	122
6.23	Accuracy Results on Pre-processed (PP) and Unprocessed (UP) With Transformer Models.	125
6.24	Pairwise p-values from Wilcoxon Signed-Rank Test for Different Feature Sets. * $0.01 < P \leq 0.05$, ** $P \leq 0.01$	127

Chapter 1

Introduction

It is fascinating what one can discover when analysing language. One interesting phenomenon of language is its social nature. Through language analysis, we draw inferences about societies. We derive insights on how people interact, the patterns of social behavior, and their cultural identities. Sociolinguistic approaches are undertaken to study the relationship between language and varying notions such as social identity and social interaction (Nguyen et al., 2016). This study studies language from a sociolinguistic perspective. A lot can be inferred about a society when linking social factors to language use.

In studying language in society and the ways in which linguistic resources and access to them are unequally distributed, sociolinguists give evidence of how patterns of linguistic variation reflect and contrast social differences. In studying responses language users have to instances of language use, they demonstrate the reality and power of affective, cognitive, and behavioural language attitudes. In analysing how language users create links between language varieties and users, institutions, or contexts, they uncover language ideologies that create social realities (Bassiouny, 2020, p.15).

Investigating language spoken or written by people from a specific geographical and cultural background provides an opportunity to extract information about the values, beliefs, cultures and identity constructions that are embedded within the words they use. The field of sociolinguistics has developed significantly with the advent of technology that has benefitted sociolinguists by offering computational techniques that make it possible to process large quantities of data (Shuy, 2003). Building upon previous insights into language and gender, this study aims at analysing the social variable of gender in the Kuwaiti Arabic (KA) dialect from a sociolinguistic perspective using computational techniques. The analysis will be conducted on social media data as it provides vast opportunities to gather large amounts of data and examine social, linguistic and cultural phenomena. Furthermore, the study will also

investigate the task of predicting gender from language use. To the best of our knowledge, no gender classification systems have been built for the KA dialect. Building such a system using a supervised machine learning approach and sociolinguistic features of KA will help in inspecting the KA dialect used in social media to explore the interplay between language, gender, and society.

1.1 Overview of Thesis Contributions

This thesis contributes to the field of Arabic Natural Language Processing, particularly addressing the gap in resources for dialectal Arabic. The main contributions of this thesis are:

1. The Kuwaiti Arabic Gender-labelled Dataset (KAGen):

This dataset is a pioneering effort in providing publicly available resources for Kuwaiti Arabic (KA) used in social media and more specifically, in the WhatsApp mobile application. KAGen consists of WhatsApp conversations that have been collected and annotated with gender labels. Its availability fills a critical void in the research community by offering opportunities to study linguistic gender representations in social media discourse.

2. The Kuwaiti Arabic Conversational Dataset (KACD):

Another contribution of this thesis is the development of a framework to annotate KAGen according to the conversational functions employed in the conversations. The framework was adopted to create the Kuwaiti Arabic Conversational Functions Dataset (KACD) which extends the analytical capabilities of KAGen by systematically categorising and annotating conversational functions observed within the dataset. This dataset is the only publicly available dataset of KA that has been annotated according to conversational functions and can be used by researchers interested in social media discourse analysis in KA.

3. The Kuwaiti Arabic Twitter Dataset (KATD):

To address the lack of open source large-scale datasets of KA, this thesis bridges the gap by providing The Kuwaiti Arabic Twitter Dataset (KATD) which is the only publicly available large-scale KA dataset labelled by gender. KATD offers researchers a valuable resource for computational processing and gender classification tasks.

Beyond dataset creation, this thesis presents three innovative studies that utilise these resources demonstrating the application of computational sociolinguistics in analysing language use patterns within the language of Kuwaiti men and women. These studies unfold in three separate chapters of this thesis:

1. **Study 1:** presented in chapter 4, explores gender-specific linguistic patterns in KA WhatsApp communication using KAGen. This study conducts both quantitative and qualitative analyses of linguistic features used by female and male participants to provide insights about gender indicative features in KA social media discourse. Additionally, it introduces a basic gender classification system, trained and tested on the KAGen dataset. Part of this work has been published in the proceedings of the Arabic Natural Language Processing Workshop (WANLP 2022), that was held in conjunction with the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022).
2. **Study 2:** presented in chapter 5, uses a thematic analysis approach to identify key conversational functions employed in KAGen. The study aims to annotate these WhatsApp conversations with their corresponding conversational function tags to create The Kuwaiti Arabic Conversational Dataset (KACD). This dataset enables further analysis of conversational interaction in Kuwaiti Arabic as used in digital platforms such as WhatsApp, with a focus on how the use of these conversational functions vary between Kuwaiti men and women in WhatsApp conversations.
3. **Study 3:** presented in chapter 3, is to the best of our knowledge the first large-scale gender classification study on KA. It involves the compilation of the Kuwaiti Arabic Twitter Dataset (KATD) that consists of tweets by KA users that have been labelled with gender labels. The study presents the development of a gender classification system that has been trained and evaluated using KATD. Two supervised learning approaches were adopted in the development of the system: a feature-engineering approach and a deep learning approach using Transformers. The study sheds lights on gendered communication patterns on Twitter through KATD, and differences in linguistic behavior of men and women in the KA Twitter community.

Overall, the aim of this thesis is twofold: to provide valuable computational datasets of KA to the Arabic Natural Language Processing community and to conduct insightful analyses into the gendered linguistic behavior observed among Kuwaiti men and women in social media contexts to identify linguistic differences between Kuwaiti men and women that may inform the development of a gender classification system. Through this approach, the thesis

offers researchers a rich resource for exploring Kuwaiti Arabic discourse in social media and enables a wide range of analyses, including language modeling, gender classification, conversation analysis, social network analysis, and cross-cultural studies.

1.2 Overview of the Thesis

The thesis is structured as follows:

- **Chapter 2:** Introduces the field of Computational Sociolinguistics and reviews relevant literature that explores the relationship between social variables, such as gender, and language usage. It provides an overview on the linguistic landscape in Kuwait and examines the interaction with social media platforms. It also discusses the challenges associated with preprocessing the Arabic language and surveys existing preprocessing tools and datasets. The chapter concludes by outlining the main task of the thesis, which is text classification, and reviewing common text classification methods.
- **Chapter 3:** Presents the research questions proposed in the thesis and outlines the methodologies adopted to address them.
- **Chapter 4:** Provides an overview of a pilot study conducted to collect a WhatsApp dataset of conversational Kuwaiti Arabic and label it with gender labels. Additionally, it conducts a quantitative and qualitative analysis of gender-indicative features to help inform building a gender classification system for KA. It then discusses the development, training, and testing of a basic gender classification system using the dataset.
- **Chapter 5:** Explores conversational functions used in the KAGen WhatsApp dataset. It introduces a framework developed to annotate the WhatsApp interactions according to the conversational functions employed. It conducts a quantitative analysis of the conversational functions used by Kuwaiti men and women, and sheds lights on qualitative observations of gendered linguistic behavior in the use of the conversational functions.
- **Chapter 6:** presents a large-scale study of gender classification of KA used in social media. It discusses the methodology followed to collect the large-scale dataset of KA tweets and annotate it with gender labels. It also details the process of developing a gender classification system for KA using feature engineering and deep learning techniques.

- **Chapter 7:** Concludes by summarising the contributions of the thesis, discussing the findings, and outlining directions for future work.

Chapter 2

Background

2.1 Computational Sociolinguistics

Language, being inherently social, embodies rich features that reflect the social context of its speakers. Sociolinguists have long been engaged in exploring the relationship between language and society, developing sociolinguistic frameworks, formulating insightful theories, and employing both quantitative and qualitative methods to study language phenomena.

Traditionally, sociolinguists have approached diverse angles, exploring topics such as language change over time within specific languages or dialects, variations in language usage across different age groups, and differences in language patterns among genders. These studies have yielded profound insights into the dynamic interplay between language and social factors. However, they often faced limitations due to small datasets or exhaustive human effort requirements.

The contemporary era is marked by fortunate technological revolutions that have revolutionised the study of these language aspects in novel and advanced ways previously unimaginable. The advent of technology and ongoing breakthroughs in Computational Linguistics (CL) have opened new avenues for collaboration between the fields of computational linguistics and sociolinguistics. This intersection has given rise to the emerging field known as Computational Sociolinguistics, where both communities can benefit from each other's expertise and methodologies.

Social variables that influence language variation include gender, age, geographical location, ethnicity, and social class. Studies in computational sociolinguistics have explored these variables and their relationship with language using computational methods. For example, there has been considerable research on automatically predicting age or gender from language using computational techniques (Bamman et al., 2014; Burger et al., 2011; Koppel et al., 2002; Suero Montero et al., 2014; Zhang et al., 2011). This is feasible due to

the availability of labeled data, where linguistic data are annotated with known values for these variables, which allows for the application of supervised learning techniques.

Researchers in the field of Computational Sociolinguistics can explore multiple sociolinguistic topics in innovative ways through this interdisciplinary approach which opens up new opportunities for studying intriguing subjects like the impact of gender on language and how language evolves over time. As technology advances, Computational Sociolinguistics will continue to deepen our understanding of how language influences and reflects social dynamics in our world.

2.2 Language and Gender

Language is a rich source for analysis and a lot of studies have been conducted to explore the relationship between different social variables and the language they construct (Eckert and McConnell-Ginet, 2013; Holmes and Meyerhoff, 2008). One of the social variables that is studied in relation to language is gender. Traditional studies of language and gender that have been conducted in the humanities and social sciences have had inconsistent findings and have received some criticism. For example, Wareing (1996) criticised the generalised insights related to the relationship between language and gender that were dependent on small samples of data. The implication from this criticism is to improve gender and language studies by using larger samples of data and different contexts (Litosseliti and Sunderland, 2002). The advent of the ‘big data’ era has revolutionised the landscape of gender analysis and presented numerous opportunities to improve language and gender studies. Moreover, investigations into sociolinguistic gender patterns have predominantly relied on qualitative methodologies, including interviews, surveys, recordings, and manual observations. However, scholars such as Bamman et al. (2014) advocate for a synthesis of qualitative and quantitative approaches, positing that while qualitative analysis sheds lights on linguistic phenomena, quantitative methods enable exploration on a broader scale and aid in the identification of cases suitable for qualitative analysis. As Litosseliti and Sunderland (2002, p.62) explain:

Language and gender may, then, legitimately be viewed from different perspectives: a pragmatic combination of methods and approaches, along with an acknowledgment of their possibilities and limitations, might allow us to focus on different aspects of the relationship between language and gender, or have a wider range of things to say about this.

2.2.1 Interdisciplinary Perspectives on Gender and Language

Renowned scholars of sociolinguistics have contributed to this prominent area of research on gender and language (Eckert and McConnell-Ginet, 2013; Holmes and Meyerhoff, 2003; Lakoff, 1973; Tannen, 1990). Lakoff (1973) is amongst the early scholars who studied the relationship between language and gender and how language reflects power dynamics. Her studies were critiqued due to depending on personal observations rather than comprehensive data analysis.

Tannen (1990) focused on differences in communication styles amongst men and women and explained the different conversational goals and strategies that influence linguistic interactions. Cameron (1998), on the other hand, questioned the common assumptions on differences in language between men and women and conveyed that language is a complex aspect of social interaction influenced by various factors other than gender such as race, social class, and cultural background. Similarly, Eckert and McConnell-Ginet (2013) argued against fixed and innate differences between men and women and view gender to be a social construct.

Language and gender studies conducted in the social sciences revealed insightful observations through qualitative methodologies (Angouri and Baxter, 2021). However, with the changes in communication platforms and continuous emergence of new social media platforms, new interdisciplinary methods are required to process large textual datasets and analyse linguistic phenomena. In addressing the challenges posed by the evolving landscape of communication platforms, computational sociolinguistics steps in to offer innovative methodologies and tools to capture the interplay between language and social interaction in the digital age.

One study conducted by Bamman et al. (2014) investigated the relationship between gender, linguistic style, and social networks. They compiled a novel corpus of 14,446 Twitter users with their gender labels (whether they are male or female) and analysed their lexical choices and their social networks using a quantitative computational approach. The approach they adopted looks beyond gender being a binary variable and explores accommodation to gender preferential language to capture how gender is a social practice. Their approach takes into consideration the “multifaceted nature of gendered language style” (Bamman et al., 2014). They first clustered Twitter users based on word counts to identify clusters of similar language styles and topics of interest. These clusters were in some cases related to a specific gender, however, there were cases that contradicted the general gender-oriented pattern noticed. Clustering was done to analyse the frequent linguistic markers in every cluster. They also trained a logistic regression classifier used for gender prediction and measured the classifier’s confidence for each user. The classifier’s accuracy was 0.88. They

were interested in looking at the cases that were not predicted correctly and performed a qualitative analysis to find explanations. Analysis of cases that defy the classifier's model for their gender revealed that social network homophily is a main factor that affects the language of the user. This means that the less same-gender people found in the user's social network, the larger the possibility of finding the user using language that is similar to the other gender.

Furthermore, a prominent research area gaining attraction within Computational Sociolinguistics is gender classification which is a task centered on predicting an individual's gender based on their language use using state-of-the-art machine learning methods. For example, Koppel et al. (2002) built a system for gender classification using machine learning algorithms on a corpus of 566 documents from the British National Corpus (BNC). They used function words and part-of-speech n-grams both separately and in combination as features and noticed that using them in tandem achieves the highest accuracy score of approximately 0.80. They also noticed that the same features performed well in classifying fiction and non-fiction texts. Findings of their research showed that women use more negation, pronouns, and prepositions such as 'for' and 'with' than men, and men use more determiners, numbers and modifiers. In another study, Zhang et al. (2011) conducted a gender classification task on an Islamic women's online political forum. They collected 17,785 female messages and 16,572 male messages and experimented with different sets of lexical features such as word length features, syntactic features such as punctuation marks and function words, structural features such as total number of sentences per message and content-specific features such as pre-selected unigrams and bigrams. The feature combination that achieved the highest accuracy was the combination of all the previous features using a support vector machine. They achieved an accuracy score of 0.86. They also noticed that the topics that women talked more about were family members, God, peace, marriage, good will, while men talked more about extremism and belief. Sec. 6.2 provides a detailed overview of gender classification studies on Arabic and English datasets.

2.3 Language and Gender Studies on the Arabic Language

"Arabic is spoken as a native language by some 250 million people in 22 separate countries collectively known as the "Arab world", which stretch from the Arabian Gulf in the east to the Atlantic Ocean in the west" (Al-Wer, 2014, p.397). In recent years, there has been an increasing interest in Arabic natural language processing topics and that can be seen for example in the NLP shared tasks that have lately included Arabic datasets for processing. However, the field is still under-researched and many research gaps form opportunities for significant contributions. Not many Arabic NLP gender studies have been carried out. The

available studies are stepping stones towards the development of this interesting research topic. Alsmearat et al. (2014) studied gender text classification of Arabic articles using the Bag-of-Words (BoW) approach. They collected and manually labelled 500 Arabic articles from different Arabic news websites. The number of articles was distributed equally across both genders. They wanted to explore the result of performing feature reduction techniques such as principal component analysis (PCA) and correlation analysis on the high-dimensional data in combination with different machine learning algorithms for the gender classification task. Results showed that Stochastic Gradient Descent (SGD), Naive Bayes Multinomial (NBM) and Support Vector Machines (SVM) were the classifiers that performed best on the original dataset where the accuracy results surpassed 0.90. Applying different feature selection techniques did reduce the dimensionality, thus improving memory and running time. However, it did not have a great effect on improving the accuracy of classifiers such as SVM, NBM, SGD and logistic classifiers. It did however significantly improve the accuracy of the K-Nearest neighbor (KNN) classifier from 0.62 to 0.82.

Shared NLP tasks that are organised for the research community have started off by tackling problems with the English language and they recently added Arabic datasets for shared tasks corresponding to the increasing interest in Arabic NLP. For example, the PAN 2017 Author Profiling Shared Task included two tasks: gender identification and language variety identification of Twitter users. Arabic, English, Portuguese, and Spanish datasets consisting of tweets were provided for training and testing. The system that achieved the highest accuracy result on gender identification of the Arabic dataset was the system developed by Basile et al. (2017). They used an SVM classifier in combination with word unigrams and character 3- to 5-grams. They achieved an accuracy result of 0.80. An interesting finding they discussed was that hand-crafted and external features such as emoji word lists, adding previous PAN data, removing certain word patterns decreased the accuracy of the system developed.

2.4 Kuwait: Demographics, Globalisation, and the Internet

Before we dive into our computational linguistic approach to studying gender in Kuwaiti Arabic, reviewing the landscape and historical background of Kuwait could help broaden our understanding and provide valuable context for our research. Kuwait is a country in the Arabian peninsula. In the early days of Kuwait, it was divided into four main areas: Sharq (the east part), Jibla (the west part), Hay Alwasat (the middle neighborhood), and Almirgab (between Sharq and Jibla). The first three areas were mainly inhabited by families

who are descendants from Najd in Saudi Arabia. Inhabitants living in these areas were known for pearl diving and trading and therefore were called 'people of the sea'. However, inhabitants of Almirgab were mainly Bedouins who lived in the South area far from the sea. Almirgab was inhabited by a lot of non-Kuwaitis, mostly immigrants who migrated from nearby countries. Almirgab was considered the poorest area in Kuwait (Al-Qenaie et al., 2011). Kuwait has a rich history that has witnessed many political, social and religious events that have had an impact on the dialects, identity, lifestyle, attitude and culture of Kuwaiti people. The British protectorate in 1899, the discovery of oil in the 1930s, Kuwait's independence from Britain in 1961, the Iraqi invasion in 1990, Kuwait's liberation in 1991 and the granting of political rights to Kuwaiti women in 2005 are some key events in the history of Kuwait.

Globalisation has significantly influenced modern Kuwait. However, Kuwait has long been governed by social traditions, cultural norms, and religious values to preserve its identity and heritage. Consequently, Kuwaiti people's attitudes toward globalisation have been in a constant state of conflict. Mahgoub (2004, p.516) explains:

Kuwait is experiencing, as in other developing countries, the tension between the forces of globalization and localization. On one hand, people are eager to enjoy the luxuries of modern life that they can afford to have while at the same time retaining a cultural identity and satisfying special social requirements.

Media globalisation and technological developments such as the internet have resulted in significant changes to the way Kuwaitis communicate in the real and virtual worlds. This is noticed particularly amongst the youth who are constructing a new identity as exposure to the western world increases. Wheeler (2000, p.443) who studied the role of new media and globalisation on the Kuwaiti national identity explains:

Internet use by youths is creating new forms of communication across gender lines, interrupting traditional social rituals, and giving young people new autonomy in how they run their lives. Although these capabilities remain tempered by pre-existing value systems, we are seeing important signs of experimentation which cannot help but stimulate processes of change over time as young people redefine norms and values for future generations.

Wheeler's observations highlight the profound impact that media globalisation and internet use have on the youth of Kuwait, particularly in their quest to balance traditional values with modern influences.

2.5 Language in Kuwait

Modern Standard Arabic is the official language of Kuwait and is the principal medium of instruction in education and formal communication, but Kuwaiti Arabic (KA) is the urban spoken dialect used in everyday informal communication. There are also other Arabic dialects spoken in Kuwait by non-Kuwaiti Arabs who have come to work in Kuwait such as Palestinian, Iraqi, Syrian, Egyptian and Lebanese (Alharbi, 1992). Kuwaiti Arabic has been influenced by the language of multi-regional immigrants and external factors such as trade. Therefore, linguistic phenomena such as lexical borrowing can be clearly noticed. There are many words in the KA lexicon that have been borrowed from languages such as English, Urdu, Turkish and Persian. However, the KA dialect has maintained a lot of its vocabulary and has not been significantly affected by external factors. For example, the word for bread in KA was and still is “khoboz”, while in some of the dialects that KA has been in contact with and influenced by call it “aish” such as Cairene Arabic and Hijazi Arabic, while in KA “aish” means rice (Al-Qenaie et al., 2011).

English is the second language of Kuwait and is widely used in workplaces, social media, and is compulsory in schools and colleges. It is the principal medium of instruction in private schools and colleges. The continuous exposure of Kuwaitis to the western world through media or academia has played a role in making the English language an influential language used by many Kuwaitis from all ages but especially the youth. The phenomenon of code-switching from KA to English is seen in informal verbal and written communication between the youth and is constantly increasing. Another language phenomenon is seen in written communication called “Arabezi” which was invented by the youth and involves using English characters and numbers to represent Arabic letters and is favored by the youth in text messaging or social media communication for different reasons such as the unavailability of the Arabic keyboard, ease of code-switching to English, or to prevent spelling or grammar mistakes in Arabic to those who are more proficient in English than Arabic (Bies et al., 2014).

2.6 Language, Gender, and Social Media in Kuwait

Kuwait is often regarded as a conservative country where cultural and religious ideologies govern the way many people think and behave. Gender segregation is a defining factor in many aspects of social and educational life in Kuwait. This segregation is evident in education, where male and female students are often taught in separate classrooms, as well as in settings such as banks, mosques, beauty salons, and most health clubs (Algharabali, 2010). The influence of gender extends beyond physical segregation and permeates interpersonal

interactions and upbringing practices in Kuwaiti families. In a typical Kuwaiti household, gatherings may take place in the same residence, but they often occur in non-mixed settings. Young boys are encouraged to participate in discussions with men in the “diwaniya”, a social gathering space where men engage in conversations about social and political matters. Conversely, young girls are encouraged to play together and assist their mothers with household chores. This distinct approach to upbringing may contribute to the formation of gendered identities through separate socialisation experiences.

Several years ago, topics related to gender, especially those involving women’s engagement in society and politics, were considered highly provocative and controversial in Kuwaiti society. However, there has been a notable shift in recent years. The advent of social media has played a major role in this shift because it made access to the online world and sharing views and opinions possible for everyone. This made opportunities for mixed-gender interactions notably increase. In the past, accessing platforms to express opinions on societal issues was limited to traditional media channels such as newspapers, television interviews, magazines, and radio broadcasts. However, these channels were often controlled by individuals with high authority or affiliation with specific political or religious groups, leading to biased perspectives. With the advent of social media, the landscape has changed dramatically. Now, not only famous public figures but also ordinary individuals have the opportunity to voice their thoughts on various matters. Social media has also contributed to destigmatising gender-related topics, as these discussions are now widespread and accessible on online platforms. Additionally, the option of interacting anonymously has encouraged women in Kuwait to participate more actively in online discourse. Social media platforms offer a promising medium for studying the language of Kuwaitis. The high level of freedom of expression enjoyed by Kuwaiti citizens allows for genuine opinions to be expressed openly, and provides ample material for sociolinguistic analysis.

Furthermore, advancements in technology have had a direct impact on communication between males and females in Kuwait. Although the general attitude towards male and female communication in public is governed by conservative cultural ideologies reflected in conscious language choice and formal interaction, online platforms have provided the opportunity to explore the nature of language used more freely amongst Kuwaiti men and women.

Sociolinguistic gender studies have been widely explored in English speaking cultures and in the Arab world too. However, there is a lack of research on the Kuwaiti dialect from a sociolinguistic perspective using large KA textual data. There are some interesting linguistic phenomena in the Kuwaiti dialect that are worth exploring. The way men and women speak is different and this can be noticed in their choice of words when for example they want to

describe something. It can be noticed that there are some words men would refrain from using because they are perceived as embodying feminine traits. For instance, Kuwaiti women use the word “eyanen” to express enthusiasm or admiration, similar to “amazing”. They might describe a movie as “eyanen”. This word is not used by men. In fact, Kuwaiti men might use “jabar” to convey enthusiasm or admiration, which in the context of describing a movie means “amazing”. Additionally, “ya hafeth” is a phrase exclusively used by women to express dissatisfaction or disappointment, translating to “Oh saver (God)”. If a man were to use this phrase, he might be perceived as speaking in a feminine manner.

Analysing the language of women and men in Kuwait offers an opportunity to explore how gendered identities are constructed and perceived. In fact, Algharabali (2010) investigated how gender representations are found in online Kuwaiti chat rooms. She also looked at how social identities are constructed in chat rooms and analysed interactional linguistic features such as ritual exchanges, formulaic expressions, humorous chat and identity symbols used in the online chat discourse. Analysis showed that cultural beliefs and norms were present in the way men and women interacted online. Moreover, it was noticed that both men and women accommodate to each other’s linguistic styles and help each other achieve mutual communication goals.

Another study was conducted by Dashti et al. (2015), who were interested in finding out if the spiral of silence theory is applicable to Kuwaiti female students’ political discourse on Twitter. The theory describes how people with views that oppose the views of the majority refrain from expressing their views due to fear of isolation (Noelle-Neumann, 1993). Dashti et al. (2015) hypothesised that Kuwaiti women are not comfortable with sharing their political views offline, but share them freely online on social media platforms like Twitter. This hypothesis was explored through self-report of Kuwaiti female students’ by using a survey method and responses showed that the spiral of silence theory is not applicable to Kuwaiti female students as they explained that they share their honest opinions both offline and online. However, gender and knowing the person they are sharing their views with were factors that affected the extent to which they share their opinions, meaning that they could refrain from sharing their opinions if the other person was a male and/or if they did not know that person (Dashti et al., 2015).

2.7 Arabic

This thesis studies Kuwaiti Arabic, a dialect of the Arabic language. To provide context about this dialect, the following lines will introduce the Arabic language and the three main forms of it: classical written Arabic, modern standard Arabic, and dialectal Arabic. Arabic is

a Semitic language spoken natively by over 200 million individuals and is considered the fourth most used language of the internet (Boudad et al., 2018; Farghaly and Shaalan, 2009).

2.7.1 Classical Written Arabic

Classical written Arabic is the form of Arabic used in the Holy Quran and the early Arabic literary texts dating back to pre-islamic times (Abdelali, 2004; Fischer, 2013). Nowadays, it is only used when reciting Quran.

2.7.2 Modern Standard Arabic

Modern standard Arabic (MSA) is the standardised form of classical Arabic. It is the form of Arabic that is used throughout the Arab world and is known as “Fus’ha” meaning “eloquent Arabic”. This form is used in newspapers, radio broadcasts, books, journals, official legal documents, reports and other formal settings. It is also the written and spoken form used and taught in academia across the Arab world (Abdelali, 2004).

2.7.3 Dialectal Arabic

Dialectal Arabic is the informal spoken language that varies greatly between different Arab regions (Elnagar et al., 2021).

It is a mixed form, with many variations, and often with a dominating influence from local languages (prior to the introduction of Arabic) or from colonial languages. Differences between the variants of spoken Arabic can be large enough to make them incomprehensible to one another (Abdelali, 2004, p.23).

Moreover, Arabic dialects are different from MSA in morphology, syntax, phonology and lexicography. Each variety is marked by locally and geographically specific features. Arabic dialects are used in informal daily communication and are used in media such as TV shows, movies and plays (Biadisy et al., 2009). They are not used in written form except in social media communication.

2.8 Arabic Datasets for Computational Processing

In recent years, the field of Arabic Natural Language Processing (ANLP) has experienced significant growth due to the ongoing developments in machine learning and deep learning techniques. The availability of textual datasets benefits the research field as it is the basis

that researchers rely on to apply computational methods to perform many useful tasks such as machine translation, sentiment analysis, named entity recognition, and other tasks. The availability of datasets is fundamental as they are used to train and evaluate various language models and applications. Compared to other languages, such as English, the availability of large-scale annotated datasets for Arabic remains relatively limited but is rapidly growing. There has been ongoing work on providing language resources for the three different varieties of Arabic, classical, modern standard and dialectal. Below we review the research efforts that addressed these varieties.

2.8.1 Classical Arabic Datasets

There have been some notable efforts in providing computational resources for Classical Arabic (Guellil et al., 2021). Sharaf and Atwell (2012a) built and annotated (QurAna) a large corpus of Quranic verses in which pronouns and their antecedents were tagged. Their corpus consists of 24,000 tagged pronouns and an ontological list of antecedent concepts that can aid information retrieval tasks. Sharaf and Atwell (2012b) have also compiled (QurSim) a large dataset in which verses of the Quran that are semantically similar were tagged. Qursim has 7600 pairs of semantically related verses and is considered a valuable resource for researchers in the field studying semantic similarity and relatedness in short texts.

Belinkov et al. (2016) conducted a study that focused on classical Arabic, particularly the language used in early historical periods. They compiled a large historical dataset of Arabic from Al-Maktaba Al-Shamela website, spanning from the 7th century to modern times. The dataset was cleaned and lemmatised to aid in semantic analysis of classical Arabic. The final dataset comprises 6,000 texts, totaling 1 billion words, with 800 million words originating from early age Arabic sources. They have made this dataset publicly available, offering significant value to the field of digital humanities.

In another study that targeted classical Arabic, Zerrouki and Balla (2017) compiled (Tashkeela), a large dataset of 75 million diacritised words which have been extracted from 97 free published online books (classic Islamic books). Their dataset is intended to automatic diacritisation systems and Arabic corpus linguistic research.

2.8.2 Modern Standard Arabic Datasets

There has been some considerable amount of work on building modern standard Arabic corpora. Abdelali et al. (2005) compiled a MSA dataset from online news articles from 10 different Arabic countries. Their dataset consists of 113 million words in MSA and was compiled to aid machine translation and information retrieval systems. Moreover, Al-Sulaiti

and Atwell (2006, 2005) compiled the Corpus of Contemporary Arabic (CCA) to support corpus linguistic studies of Arabic. They conducted a survey analysis to determine the needs of researchers working in ANLP and Arabic teaching. Their corpus consists of million words of MSA collected from online magazines, newspapers, radio broadcast transcripts, and emails. Furthermore, in an attempt to support Arabic linguistic research, Saad and Ashour (2010) compiled OSAC, an open-source corpus from diverse online websites such as BBC Arabic and CNN Arabic. This corpus covers a wide range of topics including economy, history, education, sport, and health. Similarly, Alansary and Nagi (2014) compiled the International Corpus of Arabic (ICA) that consists of 80 million words from 11 categories including social sciences, literature, humanities, sports and other categories. The texts were compiled from online newspapers, blogs and forums, and electronic books. Aichaoui et al. (2022) also compiled a large MSA spelling error corpus that was developed to aid spelling error detection and correction systems. Their corpus was compiled from different resources such as online Arabic newspapers, Aljazeera education website and almaktaba al-shamela library.

2.8.3 Dialectal Arabic Datasets

The need for dialectal Arabic datasets has been increasing with the growing interest in natural language processing tasks specific to Arabic dialects. Several studies have aimed to address this need by creating annotated datasets representing various Arabic dialects. For instance, Diab et al. (2010) annotated dialectal text extracted from online weblogs containing Egyptian, Moroccan, Levantine, and Iraqi Arabic. Suwaileh et al. (2016) compiled *ArabicWeb16* which contains 150 million Arabic web pages crawled from the Internet. It covers both dialectal and MSA text and is made publicly available and considered beneficial for information retrieval tasks. Bouamor et al. (2018) compiled the MADAR corpus which consists of sentences in the travel domain that have been translated into 25 Arabic city dialects, and the MADAR lexicon which contains 1045 entries from the same 25 Arabic cities. Furthermore, in an attempt to compile a dataset of dialectal Arabic that includes Gulf, Levantine, Egyptian dialects, Zaidan and Callison-Burch (2011) compiled a 52M-word dataset of reader comments from three online Arabic newspapers (The Arabic Online Commentary Dataset). They labelled 142,530 sentences as MSA or dialectal. Alsarsour et al. (2018) compiled the Dialectal Arabic Tweets dataset (DART) that consists of 25,000 tweets in Egyptian, Moroccan, Levantine, Gulf, and Iraqi. It is balanced across the dialectal groups and is made publicly available.

Targeting resources for the Egyptian dialect, El-Beltagy (2016) built *NileULex*, a sentiment lexicon of Egyptian and MSA words and phrases. The lexicon contains 6000 Arabic

words and phrases and was compiled over two years. As for Levantine Arabic, Jarrar et al. (2017) compiled *Curras*, a Palestinian Arabic corpus collected from various resources such as Facebook, Twitter, blogs, and stories that consists of 55960 annotated with morphological and semantic information. Maamouri et al. (2006) developed a treebank for Jordanian Arabic that consists of 26000 Jordanian words from telephone conversations. Further more, Kwaik et al. (2018) compiled *Shami*, a corpus of Levantine Arabic including Syrian, Jordanian, Palestinian and Lebanese. Their corpus consists of 117805 sentences extracted from various resources such as Twitter, blogs and stories.

Less work has been devoted to studying the Algerian dialect. Abidi et al. (2017) collected 17 million words from YouTube comments written in the Algerian Dialect. They then extracted a comparable Algerian corpus, *CALYOU*, containing aligned pairs written in both Latin script and Arabic script. As for work that addressed building resources for the Gulf dialect in the field of (ANLP), the focus has primarily been on datasets that have multiple Gulf dialects. However, studies specifically targeting individual Gulf dialects are limited in comparison. A notable Gulf Arabic dataset is *Gumar Corpus* compiled by Khalifa et al. (2016) which comprises 110 million words from 1,200 forum novels that have been annotated with sub dialect information and names of writers of the novels. Further datasets are presented in Sec. 5.2.

2.8.4 Kuwaiti Arabic Datasets for Natural Language Processing

As can be seen in Sec. 2.8, interest in compiling Arabic datasets has been increasing and the need of these datasets for natural language processing tasks is undoubted. Although some work has been devoted for modern standard Arabic and some dialectal Arabic such as Levantine and Egyptian, not much has been done for the Kuwaiti Arabic dialect.

The computational linguistic studies that targeted the Kuwaiti dialect are limited in number. Salamah and Elkhilfi (2014) compiled a dataset of 340,000 KA tweets related to the topic of ‘interrogation of ministers’ from Twitter for a sentiment classification system. They created lexicons of Kuwaiti adjectives, nouns, verbs and adverbs and used them to train a sentiment classification system. Their system achieved 0.76 precision and 0.61 recall. Furthermore, Husain et al. (2022) also worked on sentiment analysis and compiled a sentiment analysis dataset of tweets written in KA and developed a weak supervised system to automatically label the dataset with sentiment labels (positive, negative, neutral). The dataset consists of 16667 labelled tweets. However, both datasets compiled in the former studies are not publicly available and are specific to sentiment analysis. Therefore, such resources cannot be of benefit for the research field.

2.9 Challenges in Preprocessing the Arabic Language

Text preprocessing is an important step taken to prepare the textual data for any NLP task. In this step, raw data is converted to a form that is suitable for computational processing. Preprocessing helps in reducing the noise and therefore the dimensionality of the data (Haddi et al., 2013). There are many challenges in preprocessing the Arabic language due to the complex nature of the language. Some of the characteristics of the Arabic language that make preprocessing challenging are explained below:

2.9.1 The Arabic Script

The Arabic script is used in the Arabic language and other languages such as Persian and Urdu. The Arabic language has some linguistic properties that differ from the English language and make computational processing a challenging task. The Arabic language is written from right to left. In the Arabic language, the shapes of the Arabic letters differ according to their position in a sentence. For example, the letter م “meem” is written as م when it is used in the beginning of a word, and م when used in the middle of a word, and finally م when used in the end of a word. Furthermore, there is no capitalisation in the Arabic language and limited punctuation which makes building preprocessing tools for the Arabic language such as tokenisers a challenging task.

Another linguistic property related to the Arabic language is that it does not have many vowels, but has diacritics that are symbols used above or below the letters to represent short vowels. Diacritics are usually used in formal texts and dropped in informal texts. They are helpful in determining the pronunciation and meanings of words especially in cases of homographs where two or more words have the same spelling but different pronunciations and meanings. For example, the word شعر has three different meanings depending on the diacritics used: (1) شَعْر is a noun that means *poem* and (2) شَعْر is a noun that means *hair* and (3) شَعَرَ is a verb that means *felt*. Absence of diacritics, which is usually the case in most formal and informal written Arabic textual data, poses a challenge in preprocessing the data as it increases semantic ambiguity (Hanbury et al., 2011). Figure 2.1 shows the Arabic alphabet in a diacritised and undiacritised form (the main letter in the box is without diacritics and the three examples below the main letter are with diacritics).

2.9.2 Rich Morphology of the Arabic Language

Morphology is “the study of the form of words in different uses and constructions” (Matthews, 2009). The Arabic language is morphologically rich. One Arabic word could be an equivalent

ث	ت	ب	أ
ثُ ثِ ثٍ	تُ تِ تٍ	بُ بِ بٍ	أُ أِ أٍ
د	خ	ح	ج
دُ دِ دٍ	خُ خِ خٍ	حُ حِ حٍ	جُ جِ جٍ
س	ز	ر	ذ
سُ سِ سٍ	زُ زِ زٍ	رُ رِ رٍ	ذُ ذِ ذٍ
ط	ض	ص	ش
طُ طِ طٍ	ضُ ضِ ضٍ	صُ صِ صٍ	شُ شِ شٍ
ف	غ	ع	ظ
فُ فِ فٍ	غُ غِ غٍ	عُ عِ عٍ	ظُ ظِ ظٍ
م	ل	ك	ق
مُ مِ مٍ	لُ لِ لٍ	كُ كِ كٍ	قُ قِ قٍ
ي	و	هـ	ن
يُ يِ يٍ	وُ وِ وٍ	هـُ هـِ هـٍ	نُ نِ نٍ

Figure 2.1 The Arabic Alphabet with and without Diacritics. Reproduced from https://www.shamsaat.com/printables_packages

to five English words. For example, the word **وسيسألونها** as shown in Figure 2.2 is considered one Arabic word, but if translated into English is: *and they will ask her*. Each colored segment in that word has a POS and corresponds to the English word written using the same color. This shows how one Arabic word could be packed with grammatical relations between its constituent parts. This forms a challenge when building tokenisers, especially because the tokeniser's performance depends on its ability to identify the different grammatical units within an Arabic word. More specifically,

Tokenisation requires knowledge of the constraints on concatenating affixes and clitics within Arabic words. A distinction needs to be made between clitics which are syntactic units and thus have their own part of speech and affixes that mark grammatical inflections such as tense, number and person agreement (Farghaly and Shaalan, 2009, p.12).

Another major challenge in processing the Arabic language is homographs where one word could belong to more than one part of speech. Diacritics may help in this case such as it has been explained in the previous section, however the challenge remains when performing normalisation which is recommended to deal with the variability in the Arabic script. Another challenge that can be faced in segmenting Arabic text is when a word can be segmented in different forms (Farghaly and Shaalan, 2009). This can happen in both MSA and dialectal Arabic. For example, the word **النزهة** could be a proper noun that refers to the name of a city in Kuwait or to a noun preceded by the definite article **ال**, which means “the trip”. An efficient segmenter should be able to identify that in the former case the whole word is one segment and in the later case the word should be segmented into two segments: the definite article and the noun. One way of dealing with these challenges is using a morphological analyser and disambiguator. In fact, Obeid et al. (2020) built a morphological analyser that predicts all possible morphological features for a given word and a disambiguator that ranks the generated out-of-context analyses of the word and chooses the top ranked analysis depending on pre-computed probabilistic score of both the lemma and POS frequency. These two steps help in many Arabic NLP tasks by offering more accurate morphological representation of words.

وسيسألونها
and they will ask her

Figure 2.2 An Example of One Arabic Word and its' English Equivalent Represented by the Same Color

The former points discussed address some of the challenges that are faced in preprocessing the Arabic language. As can be seen, the lack of punctuation and diacritics, the absence of capitalisation, the inconsistency in spelling letters and words which directly affects their meanings, the morphological ambiguity in Arabic such as that seen in cases of homographs are all challenges that make preprocessing a difficult task. However, there are some proposed

solutions such as using a morphological analyser and disambiguator to aid preprocessing techniques such as normalisation, segmentation and POS tagging (Khalifa et al., 2020; Obeid et al., 2020).

2.10 Arabic Preprocessing Tools

As interest in Arabic natural language processing has been increasing in recent years, efforts have been made to build and improve tools for processing the Arabic language. Processing tools for the Arabic language are scarce compared to tools developed for Western languages. However, there are some tools that have been built for the Arabic language and are still undergoing improvement to deal with the challenging issues of the language such as those raised in the previous section. Preprocessing tools that support MSA perform better than tools built to process dialectal Arabic due the availability of large MSA corpora as opposed to dialectal Arabic corpora and also because of the higher complexity and richer variety of Arabic dialects. This has made building dialectal tools an exceedingly challenging task and only few tools that support dialectal Arabic data are available. Below is a list of some available Arabic preprocessing tools that have been tested and are compared in Sec. 2.11 in order to select which tools to use for the experiments reported below.

Stanford CoreNLP Stanford CoreNLP is a natural language processing toolkit that enables users to apply preprocessing techniques such as POS tagging, named-entity recognition, tokenisation, dependency parsing, sentence splitting and other language processing techniques. Stanford CoreNLP was designed to process the English language but has models that support other languages such as Arabic, Spanish, Chinese, French and German. All of the features can be used on the English language, however, not all features are applicable to the other languages. As for the Arabic language model, it offers tokenisation, sentence splitting, POS tagging, and constituency parsing (Manning et al., 2014).

The Classical Language Toolkit (CLTK) CLTK is another natural language processing toolkit that offers models to process different languages. It supports the Arabic language in tokenisation. It also uses Pyarabic which is a Python library that offers preprocessing techniques such as tokenisation, removing diacritics, extracting numerical phrases and other techniques (Johnson et al., 2014–2020).

MADAMIRA MADAMIRA is a morphological analysis and disambiguation tool that has been created exclusively for the Arabic language. It has been built using machine learning

algorithms and a morphological analyser to process MSA and lately Egyptian Arabic. This tool performs preprocessing techniques such as tokenisation, diacritisation, POS tagging, named entity recognition and other techniques (Pasha et al., 2014).

FARASA FARASA is a fast Arabic language processing tool that performs segmentation, POS tagging, dependency parsing, diacritisation, lemmatisation, and named entity recognition (Abdelali et al., 2016).

Tashaphyne Tashaphyne is a light Arabic stemmer. It's a Python library that performs stemming and root extraction. It has an integrated list of Arabic prefixes and suffixes and allows users to customise prefixes and suffixes by adding their own as a list. This feature is useful when segmenting dialectal Arabic (Zerrouki, 2010).

CAMel Tools CAMel tools is a set of Arabic preprocessing tools that is written in Python and is open-source. It supports MSA and dialectal Arabic. It offers different NLP utilities such as morphological modeling, sentiment analysis, named entity recognition, dialect identification and preprocessing tools such as tokenisers, POS taggers, diacritisers. Building this set of tools has been inspired by other tools such as MADAMIRA and FARASA. It therefore has promising NLP utilities that support MSA and dialectal Arabic and can be further enhanced by researchers working on the Arabic language (Obeid et al., 2020).

2.11 Comparative Analysis of Existing Arabic Preprocessing Tools

Preprocessing textual data is an important stage that is taken after the data collection phase to prepare the data for computational processing. There are many efficient preprocessing tools for the English language that are used in many natural language processing tasks. However, preprocessing tools for the Arabic language are limited and work has been done for MSA but not much work has been done on dialectal Arabic. Section 2.10 reviewed the available preprocessing tools for the Arabic language. The following subsections tests and compares the existing preprocessing tools for the Arabic language to determine which tools will be used in the experiments of this thesis.

The preprocessing tools have been tested on two different Arabic social media textual datasets compiled from two different social media platforms, Twitter and WhatsApp, to determine their effectiveness. The Twitter dataset that has been used is from the PAN 2017 Author Profiling Shared Task (Rangel et al., 2017). Permission to access the Arabic dataset

and use it for research purposes has been granted by the Shared Task organisers. The Gulf dialect dataset has been downloaded to be used for preprocessing because it includes KA along with other Gulf dialects that are very similar. The dataset consists of 60,000 tweets. The WhatsApp dataset has been compiled from one of the WhatsApp groups participating in our first experiment and consists of 1028 sentences. The preprocessing tools mentioned in Section. 2.10 have been tested on the compiled Twitter and WhatsApp datasets. The experiment involved testing tokenisation, stop-word removal, diacritisation and POS tagging on both MSA and dialectal Arabic. A comparative analysis of the preprocessing tools is provided in the following.

2.11.1 Tokenisation, Removing Diacritics, and Stop-word Removal

Applying tokenisation to the Arabic language is not an easy task due to the rich and complex morphology of Arabic. Tokenisation for the Arabic language is “the division of a word into clusters of consecutive morphemes, one of which typically corresponds to the word stem, usually including inflectional morphemes” (Habash, 2010, p.66). Because one Arabic word may consist of multiple morphemes that belong to different parts of speech, it can be challenging to determine the boundaries within a word. One Arabic word could consist of a noun or verb conjugated with a pronoun, preposition, or conjunction. Habash (2010) explains that there is not one correct way to segment an Arabic text, but rather the process depends on what morphemes the researcher is interested in segmenting.

The preprocessing tools mentioned in Sec. 2.10 were tested on the Twitter and WhatsApp datasets discussed at the outset of this section with the aim of testing how well they perform in terms of runtime and accuracy of outputs. The tools were tested on a subset of 5000 tweets from the Twitter dataset. Stanford CoreNLP tokeniser took 44.52 sec to tokenise the tweets, CLTK took 0.08 sec and CAMEL took 0.03 sec. As for the WhatsApp messages, the tools were tested on 1028 messages and Stanford CoreNLP tokeniser took 7.95 sec, CLTK tokeniser took .007 sec, and CAMEL took 0.01 sec. CLTK tokeniser and CAMEL tokeniser performed incredibly well in terms of runtime.

Samples from Twitter and WhatsApp messages were extracted for manual inspection and it was noticed that the Stanford CoreNLP tokeniser did generally well in tokenising the sentences and clitics. It however, was not able to identify emojis and replaced them with question marks during tokenisation. CLTK, on the other hand, did not segment clitics but was able to identify and segment emojis. CAMEL performed well in tokenisation but deleted emojis when tokenising.

Fig. 2.3 shows the outputs of the Arabic tokenisers. The underlined segments are morphemes attached to the words that refer to articles or pronouns.

The sentence: **كتبت الطالبة مقال يتحدث عنهم** that translates to: “The student wrote an essay that talks about them” was used to test tokenisation.

Tokeniser	Sentence After Tokenisation
Stanford CoreNLP	كتبت الطالبة مقال يتحدث عن+هم
CLTK	كتبت الطالبة مقال يتحدث عنهم
MADAMIRA	كتبت ال+طالبة مقال يتحدث عن+هم
FARASA	كتب+ت ال+طالب+ة مقال يتحدث عن+هم
CAMeL	كتبت الطالبة مقال يتحدث عنهم

Figure 2.3 Comparison between the Output of Three Arabic Tokenisers

It was noticed that CLTK was not able to segment the pronouns attached to the words. CAMeL has a simple tokeniser that does not segment clitics and a morphological tokeniser that is able to segment clitics. FARASA and MADAMIRA were able to segment articles and pronouns attached to the words, however FARASA was not as accurate as MADAMIRA. Stanford CoreNLP was able to segment one pronoun attached to a word, however, there were more pronouns and articles that were not segmented. This shows that MADAMIRA, FARASA, and CAMeL are better able to deal with more morphological features of Arabic than the other tools tested.

CLTK and CAMeL supported stop-word removal and diacritisation. For example, they were able to convert **اللُّغَةُ الْعَرَبِيَّةُ** which is the diacritised form of “the Arabic Language” to **اللغة العربية**, removing diacritics from the words. MADAMIRA and FARASA were able to add diacritics to undiacritised text but cannot remove diacritics from text.

2.11.2 POS Tagging

POS tagging was tested with the aim of evaluating runtime and accuracy of the output. A subset of 5000 tweets was extracted from the compiled Twitter dataset to compare runtime of the POS taggers. Stanford CoreNLP took 49.93 sec to tag the tweets, CAMeL took 16.32 sec and Farasa took more than 8 minutes to tag 100 tweets. As Farasa was very slow, it will be excluded. As for MADAMIRA, we tried their online tool on their website. Table. 2.1 provides a comparison of the output of the four different POS taggers on an example sentence from an online Arabic newspaper that is constituted of a variety of parts of speech.

The sentence translates to: **Corona virus is spreading in more than 200 countries around the world, after it started spreading in China.** The outputs of the systems were very similar and mostly accurate, however, an interesting finding was that the word “Corona” **كورونا** was not tagged correctly when using MADAMIRA and CAMeL. One possible

Segment	Translation	Stanford CoreNLP	FARASA	MADAMIRA	CAMeL
ينتشر	is spreading	Verb	Verb	Verb	Verb
فيروس	Virus	Noun	Noun	Noun	Noun
كورونا	Corona	NNP	NOUN	Verb	Verb
في	in	IN	PREP	Preposition	PREP
أكثر	more	JJR	ADJ	Comp. adj	ADJ
من	than	IN	PREP	Preposition	PREP
200		CD	NUM	Digit	Digit
دولة	country	NN	NOUN	Noun	Noun
شتى	various	NN	Noun	Noun	Noun
أرجاء	over	NN	NOUN	Noun	Noun
العالم	the world	DTNN	NOUN	Noun	Noun
،		PUNC	PUNC	PUNC	PUNC
بعد	after	NN	NOUN	Noun	Noun
أن	that	IN	PART	Sub. Conj.	Sub. Conj.
بدأ	started	VBD	V	Verb	Verb
تفشي	spreading	NN	Noun	Noun	Noun
هـ	it	PRP\$	—	—	abbrev
الصين	China	DTNNP	NOUN	Proper Noun	Proper Noun

Table 2.1 Comparison between the Output of Four Different Arabic POS Taggers for the sentence that translates to: “Corona Virus is spreading in more than 200 countries around the world, after it started spreading in China.”

explanation is that in many cases **نا** is found as a suffix in words and this suffix is usually a pronoun attached to a verb which in this case makes the whole word a verb. However, in the case where it is used to refer to the pandemic, it is not a suffix and the word has been translated literally from English to Arabic. Therefore, the word should be tagged as a noun not a verb. The system may have not seen many cases such as this to be able to infer this conclusion. However, MADAMIRA gave a more detailed description of parts of speech such as comparative adjective instead of adjective, subordinating conjunction instead of conjunction and proper noun instead of only noun.

The Stanford CoreNLP tool was the only tool that was able to detect clitics and break the word **تفشييه** into **تفشي** and **ه** and tag these segments correctly. CAMEL has a morphological disambiguator tool that provides tags of different features for given words such as POS tags, diacritised forms of words and lemmas of words. The POS tagger supports MSA and dialectal Arabic including Gulf Arabic.

2.11.3 Summary or Analysis

As can be seen, there are some useful tools available for preprocessing the Arabic language that mostly support MSA. However, there are some attempts to improve these tools to support dialectal Arabic such as MADAMIRA and CAMEL. The overall results of comparing the available preprocessing tools showed that CLTK and CAMEL performed well in tokenisation in terms of runtime, while Stanford CoreNLP showed lower performance.

In terms of POS tagging, Stanford CoreNLP, CAMEL, and MADAMIRA were evaluated for runtime and accuracy. While Stanford CoreNLP showed slower runtime, it was the only tool able to detect clitics accurately. CAMEL, on the other hand, provided detailed morphological disambiguation. We decided to use CAMEL for the experiments due to it being open-source, supporting Gulf dialects and has many methods to analyse linguistic features of Arabic such as morphological disambiguation and preprocessing utilities.

2.12 Text Classification

Text classification is a widely recognised task in the field of Natural Language Processing, involving the categorisation of text into predefined classes or categories using supervised machine learning models (Jindal et al., 2015). Text classification techniques are crucial in various real-life applications due to their capability in analysing vast amounts of data and categorising them using state-of-the-art computational tools. Such applications are beneficial to many text mining tasks, including sentiment analysis, where classification techniques are

used to classify segments of text according to their polarity labels. Additionally, they play an instrumental role in filtering out spam emails, detecting fake news, and identifying topics or themes within large text corpora (Dalal and Zaveri, 2011).

Many modern approaches to automatic text classification involve supervised learning, where texts labelled with the categories to distinguish are provided for training. The approaches divide into two groups: (1) those where text features to be used in learning a model are manually defined by a human using intuitions about which features are likely to prove useful in classification and (2) those where only the raw data, possibly transformed into another representation (e.g. by using word embeddings for each word in the input text), are used and the classifier itself learns the representations (i.e. features) it deems useful in classifying text (Sebastiani, 2002).

This thesis conducts automatic text classification to investigate whether we can classify texts by the gender of the author. More specifically, we examine the extent to which the gender of Kuwaiti social media users can be predicted from their social media textual data. Automatic text classification also provides an opportunity to test how well gender indicative features, elicited from the literature or observed within our own research, perform in accurately determining the gender of the author based on their writing style and linguistic patterns. We experiment with a feature engineering approach and a deep learning approach using Transformers.

Figure 2.4 illustrates the typical steps followed in a text classification pipeline. The process begins with the conversion of raw text into structured text through feature extraction. During this step, features are extracted from the raw text to train a machine learning classifier. Subsequently, a classifier is selected and trained using the extracted features. The final stage involves using the trained classifier to make predictions on unseen text data (test data) and evaluating the accuracy of its predicted labels through evaluation metrics (Kowsari et al., 2019).

2.12.1 Statistical Approaches with Feature Engineering

Statistical approaches to text classification often involve manual feature engineering, where relevant features are specified by a human feature engineer to represent key text characteristics and feature extractors are written to automatically extract these features from an input text. Various features such as bag-of-words, Term Frequency-Inverse Document Frequency (TF-IDF), and n-grams are commonly used in building a feature vector representation of a raw text. These engineered features are then fed into traditional machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), or logistic regression for classification.

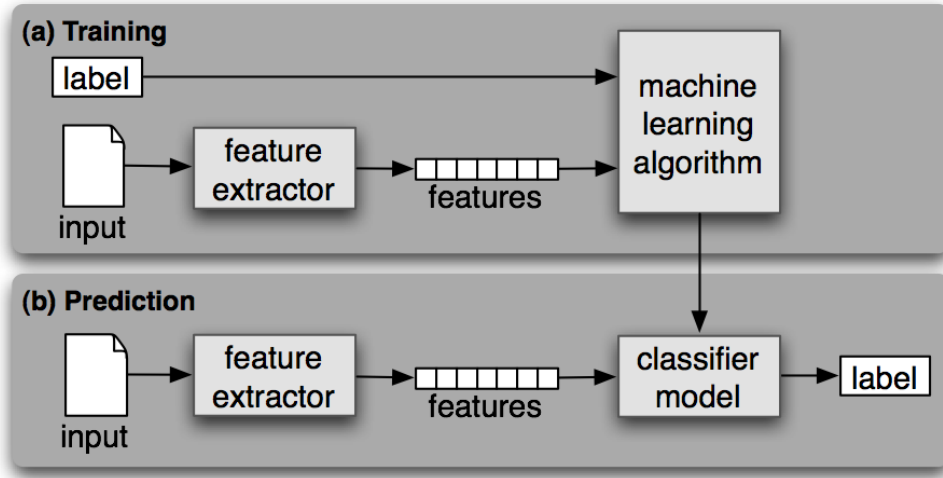


Figure 2.4 Text Classification Process. Figure from Bird et al. (2009)

Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are widely adopted in machine learning for text classification tasks, where they have demonstrated robust performance, particularly in binary classification scenarios. The primary goal when employing an SVM is to transform the original data points from the input space to a higher-dimensional space and identify an optimal hyperplane or decision boundary that effectively separates the data points into two distinct groups or classes. Fig. 2.5 shows the mapping of data points from the input space to a higher dimensional feature space to achieve linear separation (Luts et al., 2010).

This decision boundary is determined by the equation:

$$t_{new} = \mathbf{w}^t \mathbf{x}_{new} + b$$

Here \mathbf{w} is the weight vector that defines the orientation of the hyperplane, b is the bias and \mathbf{x}_{new} denotes the feature vector of a data point \mathbf{x} after it has been transformed into the new, higher dimensional feature space. The value of t_{new} indicates the position of the data point with respect to the decision boundary: positive values signify one class, while negative values signify the other. During the training phase of the SVM algorithm, the parameters \mathbf{w} and b are optimised to maximise the margin between the hyperplane and the closest data points, known as support vectors. This optimisation process ensures that the decision boundary effectively separates the two classes while minimizing classification errors (Campbell and Ying, 2022; Rogers and Girolami, 2016) as can be seen in Fig. 2.6. SVMs use Kernels to capture complex relationships between data points that may not be linearly separable in the original feature space.

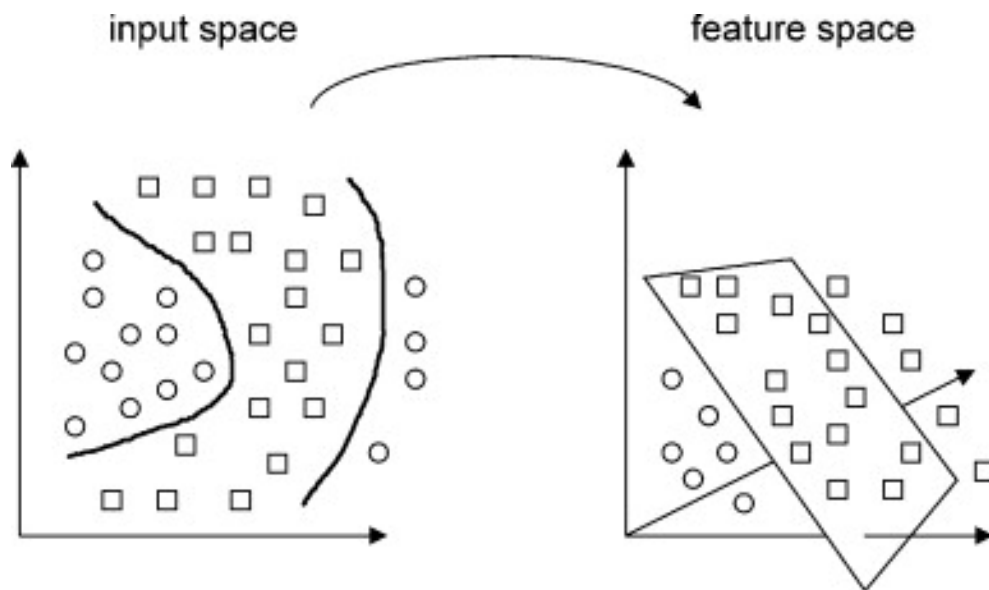


Figure 2.5 SVM mapping the data points from the input space to a higher dimensional feature space to linearly separate the data. Figure from Luts et al. (2010).

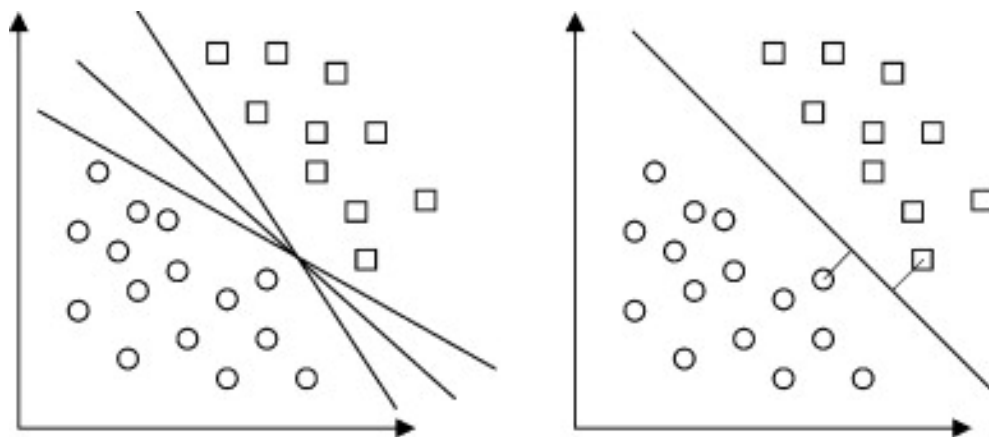


Figure 2.6 The left figure illustrates various hyperplanes capable of separating the data, while the right figure demonstrates how SVMs construct a hyperplane that maximizes the margin between the different classes of data. Figure from Luts et al. (2010).

Each feature value x is multiplied by its corresponding weight w , which indicates the feature's importance.

Logistic Regression

Logistic regression is a supervised machine learning algorithm that is commonly used for binary text classification. It is a statistical method that is based on finding the probability of

an instance belonging to one of two possible categories/classes $y = 1$ or $y = 0$. To achieve this, the classifier is trained using the weights w and the bias b . Each feature value x is multiplied by its corresponding weight w , which indicates the feature's importance. After all feature values are multiplied by their weights, they are summed up, and then the bias b is added. This is given in equation 2.1 as presented in Jurafsky and Martin (2023). Probabilities are values between 0 and 1 and this equation would give a value in the range $-\infty$ to ∞ .

However, Logistic regression uses the sigmoid function to generate probability values ranging from 0 to 1 based on the input independent variables, which is given in equation 2.2, so to ensure it represents a probability, we need to verify that the two cases, $p(y = 1)$ and $p(y = 0)$, add up to 1, this is illustrated in equation 2.3.

$$z = \left(\sum_{i=1}^n w_i x_i \right) + b \quad (2.1)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + \exp(-z)} \quad (2.2)$$

$$\begin{aligned} P(y = 1) &= \sigma(\mathbf{w} \cdot \mathbf{x} + b) \\ &= \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))} \\ P(y = 0) &= 1 - \sigma(\mathbf{w} \cdot \mathbf{x} + b) \\ &= 1 - \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))} \\ &= \frac{\exp(-(\mathbf{w} \cdot \mathbf{x} + b))}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))} \end{aligned} \quad (2.3)$$

Therefore, given two classes, Class 0 and Class 1, if the logistic function output is greater than the threshold value (a decision boundary), the instance is categorised into Class 1; otherwise, it is categorised into Class 0.

The parameters (coefficients for the features) in a Logistic regression classifier are estimated using Maximum likelihood estimation (MLE). The goal of MLE is to find the parameter values that maximise the likelihood of observing the given data under the assumed logistic regression model (Jurafsky and Martin, 2023).

There are three types of Logistic regression classifiers (Hassan et al., 2022):

1. Binomial:

used for binary classification tasks, where the outcome variable has only two possible

categories or classes. It models the probability of an instance belonging to one of the two categories.

2. Multinomial:

used when the outcome variable has more than two categories. It models the probabilities of an instance belonging to each of the multiple categories simultaneously.

3. Ordinal:

used when the outcome variable is ordered or ordinal such as in customer satisfactory ratings (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied).

K-Nearest Neighbours (KNNs)

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm that is primarily used for classification and regression tasks. Its application in text classification tasks has been widespread due to its simplicity and intuitive nature. The KNN algorithm functions by selecting a value for K , which represents the number of nearest neighbors to consider, and employing distance metrics like Euclidean distance to measure the distance between the new input instance and its K nearest neighbors (Hassan et al., 2022; Telnoni et al., 2019).

Choosing the value of k is a crucial step. Fig. 2.7 illustrates the K-nearest neighbors (KNN) decision rule in two scenarios: (a) when $K = 1$ and (b) when $K = 4$, applied to a dataset divided into two classes.

In (a), if the closest data point to the new input instance (?) is a red square, then according to the K-nearest neighbors (KNN) classification with $K=1$, the new input instance (?) would be classified as belonging to the class represented by the red squares in the dataset.

In (b), with $K=4$, the four nearest data points consist of three red squares and one green triangle. According to the K-nearest neighbors (KNN) algorithm, the majority votes (three red squares) will assign the class red square to the new data instance (?) (Imandoust et al., 2013).

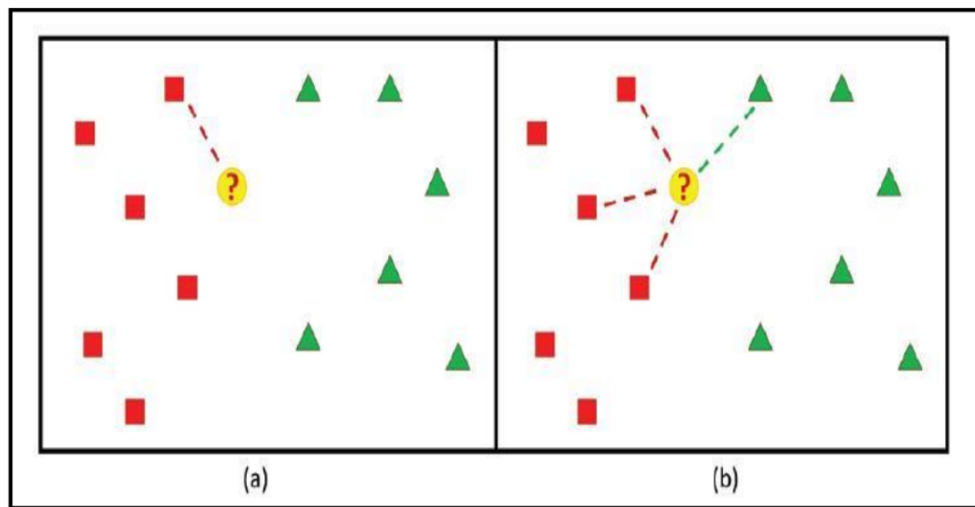


Figure 2.7 KNN Classifier from Imandoust et al. (2013)

Fig. 2.8 also shows how with increasing the value of k , the decision boundaries become smoother. As k increases from 1 to 7, the decision boundaries become smoother. This means that with larger k values, the boundaries between different classes in the dataset become less jagged and more generalized. This smoothing effect helps reduce overfitting to individual data points and allows for better generalization to new, unseen data (Moldagulova and Sulaiman, 2017).

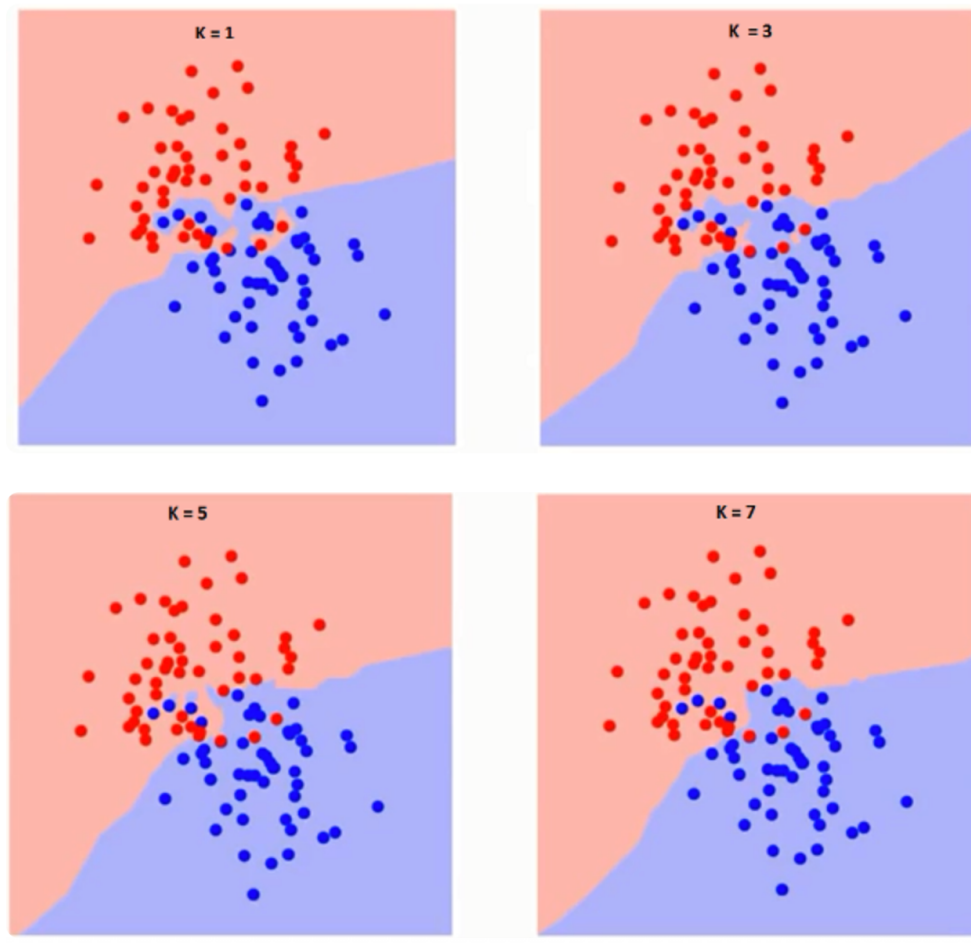


Figure 2.8 Setting K to Different Values. Figure from <https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/>

Decision Trees

Decision trees are one of the earliest and most widely used classification algorithms (Quinlan, 1986). They have some advantages that make them preferable in classification tasks such as their speed in both learning from data and making predictions. They are also highly interpretable and can be translated into a set of if-then rules which makes them suitable for rule induction systems. However, they do have drawbacks such as their tendency to overfit the training data, especially when dealing with complex datasets or noisy data (Kowsari et al., 2019).

Decision trees have a hierarchical tree structure (Charbuty and Abdulazeez, 2021). The concept behind decision trees involves breaking down classification tasks into a series of decisions about each feature, starting from the root of the tree and proceeding to the

leaves, where the final classification decision is made as illustrated in Fig. 2.9. In terms of optimisation and search, decision trees employ a greedy heuristic approach, evaluating available options at each stage of learning and selecting the one that appears most optimal at that particular point. The most optimal option is typically calculated using a measure like information gain (Quinlan, 1986). This heuristic strategy often produces effective results across a wide range of applications. (Marsland, 2011).

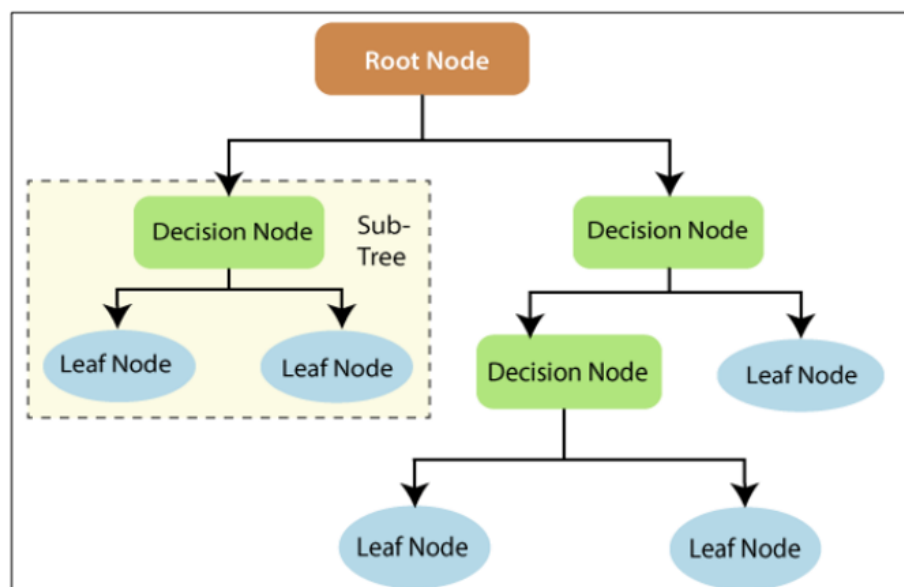


Figure 2.9 Structure of a Decision Tree. Figure from Charbuty and Abdulazeez (2021)

2.12.2 Deep Learning Approaches with Transformers

The emergence of deep learning models has revolutionised various domains within natural language processing, including text classification. One notable advantage of this approach is the absence of the need for manual feature extraction, which in a supervised feature-engineering approach involves (1) the specification of what the features are (which requires intuition on the part of the designer) (2) for each feature, the writing of the code that will automatically extract the value of the feature from an input text. Then, the actual feature extraction is done automatically. The issues are that the identification of features may be sub-optimal (a human may not be able to see what the best features are) and the writing of feature extractors takes time. Deep learning methods learn the features (these may not be human interpretable) from data and the learned model does feature extraction itself, avoiding the need to write feature extractors which is commonly viewed as a time-consuming process (Gasparetto et al., 2022). Furthermore, this approach can better capture the semantic meaning

of words and generate contextualised representations for each word within a sentence. Additionally, these models have the ability to transfer learned knowledge across similar tasks.

Transformers

Transformer models have created breakthroughs in the field of Natural Language Processing, especially in tasks involving understanding and generation of text. These models have brought about significant changes and advancements in various NLP tasks such as language translation, text summarisation, sentiment analysis, named entity recognition, question answering, and more.

Transformers were developed by a group from Google in 2018 who published the paper “Attention is all You Need” introducing the concept and architecture of transformer models that were proposed to solve various sequence transduction tasks. While it was evaluated on machine translation, its applicability extends to other tasks as well (Vaswani et al., 2017). While the area of Transformer models is vast and continually expanding, this thesis will focus on Bidirectional Encoder Representations from Transformers (BERT), as the model used in our study is a fine-tuned Arabic BERT-based model.

The transformative advancements in transformer models can be primarily attributed to the key innovation of *self-attention mechanism*, which is a mechanism employed in the Transformers architecture to learn relationships between words. This mechanism allows the model to capture dependencies between different words in the input sequence and this is done by attending to relevant parts of the input sequence simultaneously, rather than processing words sequentially (Gasparetto et al., 2022). Fig. 2.10 illustrates the information flow in a self attention layer, demonstrating how each input is attended to in relation to all other inputs in a parallel manner (Jurafsky and Martin, 2023).

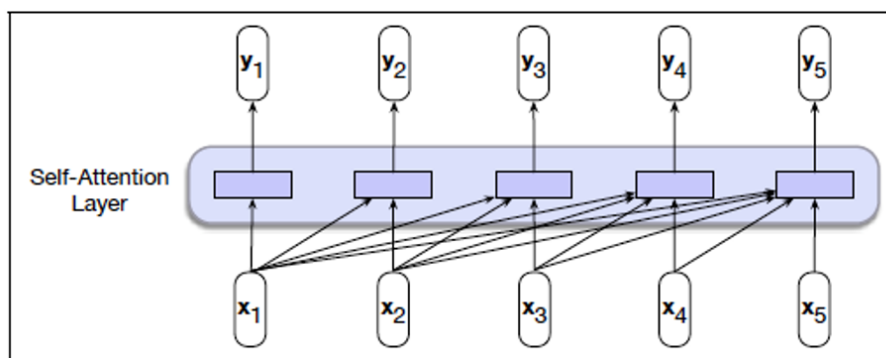


Figure 2.10 Information Flow in a Self-attention Model. Figure from Jurafsky and Martin (2023).

Jurafsky and Martin (2023) explain that the input embeddings undergo a transformation that assigns them three roles in representing the words of a sequence: query, key, and value. The query represents the current focus of attention during comparisons with preceding inputs. The key represents the preceding input that is being compared to the current input. Finally, the value plays the role of computing the output of the current input.

In a self-attention layer, the dot product is commonly used to compare the query and key embeddings. This results in a score for each input element. A higher score indicates greater similarity between the query and key embeddings. These scores are then normalised using the softmax function to obtain attention weights, which determine the importance of each input in relation to the element that is currently attended to (Jurafsky and Martin, 2023). Furthermore, *Positional encoding* helps the model identify the order of words, thereby capturing the sequential order of words (Vaswani et al., 2017).

Multi-head Attention (MHA) builds on the self-attention mechanism by using multiple attention heads. While self-attention calculates the importance of each word relative to others in a sequence, MHA takes this further by having several self-attention layers (heads) operate in parallel (Jurafsky and Martin, 2023).

Overall, the output of the encoder (vector representations for each word), is transmitted to the decoder alongside the vector representations for the target input sequence. Positional encoding is applied to both sets of representations to retain positional information. Subsequently, Multi-head Attention layers are extracted and processed through a FeedForward neural network. A linear transformation is then applied, followed by a softmax operation that assigns probabilities to each word. The word with the highest probability is selected as the output of the transformer model. This process iterates for each word in the sequence (Gasparetto et al., 2022; Tunstall et al., 2022; Vaswani et al., 2017).

Bidirectional Encoder Representations from Transformers (BERT)

In 2018, the introduction of Bidirectional Encoder Representations from Transformers (BERT) significantly advanced the capabilities of Large Language Models (LLM) in the field of NLP. BERT is considered the state-of-the-art model for many NLP tasks such as question answering, text classification, language inference and other tasks (Devlin et al., 2018b).

BERT's architecture is influenced by Transformer models. However, unlike the original Transformer, which consists of an encoder and decoder for sequence-to-sequence tasks like machine translation, BERT consists of stacked encoders only. These deep bidirectional encoders, as can be seen in Fig. 2.11, process input sequences from both directions, generating contextual representations for every word in the input sequence (Devlin et al., 2018b).

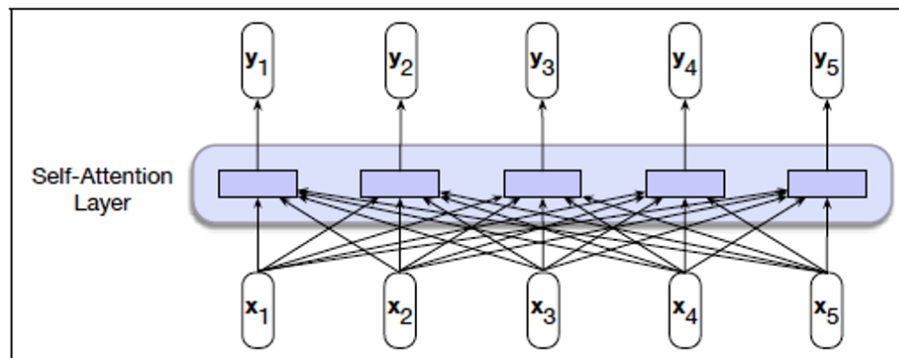


Figure 2.11 Bidirectional information flow in a self-attention model. Figure from Jurafsky and Martin (2023).

- **Pre-training BERT:**

BERT was pre-trained on two tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). In MLM, certain words amongst a sequence of words are masked and the task is to predict the masked words from the neighbouring words. For example, the model could be given a sentence: “She <mask> the gym <mask> Sunday”, and the aim would be to predict the masked words ‘joined’ and ‘on’ from the original sentence: “She joined the gym on Sunday”.

The second task that BERT was trained on is NSP in which two sentences are given and the task is to predict if the sentences are consecutive or not (Géron, 2022). For instance, when presented with the sentences “I am busy now” and “I can meet you at 3:00”, BERT should predict that the latter sentence follows the former i.e. they are consecutive. Conversely, if presented with “I am busy now” and “The cat is sleeping”, BERT should recognise that they are not consecutive. This task improves BERT’s performance by enabling it to learn relationships between sentences (Devlin et al., 2018b).

The datasets that were used to pre-train BERT are the BooksCorpus containing 800 million words (Zhu et al., 2015), and English Wikipedia consisting of 2,500 million words. When extracting text from Wikipedia, they extracted text passages only and disregarded lists, tables, and headers.

- **Fine-tuning BERT:**

BERT can be fine-tuned for many NLP tasks such as sentiment analysis, question answering, and other tasks. During the fine-tuning process, input sequences are converted into sub-tokens, and special tokens, [CLS] and [SEP], are added based on the task requirements. [CLS] is added to the beginning of the input sequence and [SEP]

is added to the end of each sequence. The [CLS] token serves as a marker to denote the start of the input sequence, while the [SEP] token indicates the separation between sentences or segments within the sequence and its placement depends on the specific design of the task. For instance, it is often added between two sentences in sentence pair tasks but may not be necessary at the end of a single sequence. Additionally, a single task-specific layer is added to the pretrained BERT model, replacing the original language modeling head. The output vector from the final layer of the [CLS] token is fed into the classifier head to make predictions (Jurafsky and Martin, 2023). For example, in text classification tasks, such as sentiment analysis, the final layer of the [CLS] token is fed into a classifier head to predict target labels positive, negative or neutral. In chapter 6, we fine-tune Arabic BERT-based Transformer models to predict the gender of Kuwaiti social media users from their language usage. Fine-tuning BERT is considered a highly efficient optimisation technique due to its reduced computational expense. (Maslej-Krešňáková et al., 2020; Vaswani et al., 2017). Fig. 2.12, represents an example of a sequence classification task using BERT. Details of the fine-tuning process conducted on our data is presented in Sec. 6.6.

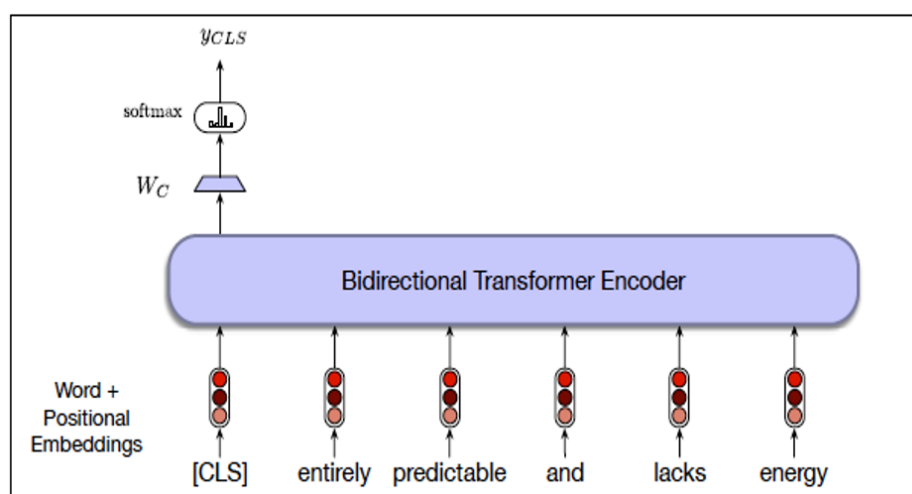


Figure 2.12 Sequence Classification Using BERT. Figure from Jurafsky and Martin (2023).

2.13 Summary

In conclusion, this chapter has introduced the field of Computational Sociolinguistics and explained how this emerging interdisciplinary field serves as a bridge between sociolinguistics and computational linguistics. It reviewed gender related work of sociolinguists, and

computational linguists and studies that incorporated both approaches. It then provided an overview about the context of the dialect this thesis focuses on by introducing the history of Kuwait, the people of Kuwait, the KA dialect and linguistic phenomena found in the dialect. It also shed light on the role of social media in Kuwait and how that social media platforms are promising platforms for linguistic analysis of sociolinguistic phenomena. It discussed the challenges in preprocessing the Arabic language and presented a comparison of available open-source tools built for Arabic preprocessing and justified the choice of tools to be used. It finally presented a background about the key computational techniques of automatic text classification that is employed in the empirical studies of this thesis.

Chapter 3

Research Methods and Data

3.1 Research Questions

This study aims to explore linguistic phenomena embedded in the KA dialect through an interdisciplinary lens. By incorporating sociolinguistic theory and computational linguistic methods, the goal is to study gender-indicative features within online language usage from multiple perspectives. The selection of this research topic and the formulation of research questions are motivated by the interest in utilising the availability of large Kuwaiti Arabic (KA) data on social media platforms.

The abundance of KA social media data presents a unique opportunity to collect and annotate datasets specific to Kuwaiti Arabic. This, in turn, helps to address the current scarcity of research in the field of Computational Sociolinguistics, particularly concerning gender analysis in Kuwaiti Arabic. Through this research project, we aim to contribute to the thriving field of Arabic Natural Processing with a specific focus on enriching our understanding of Kuwaiti Arabic used in online platforms. The following are the proposed research questions:

1. **What distinguishes the language use of Kuwaiti female and male social media users?**

This question will be answered by conducting two studies: a WhatsApp study and a Twitter study. This question aims at exploring different elements of language use between men and women such as: interactional features, emoji usage, POS usage, code-switching, words or expressions that are exclusive to each gender to find patterns that describe the language of male and female Kuwaiti users.

2. **Are there specific conversational strategies employed by Kuwaiti male and female users in WhatsApp exchanges and do they vary between both gender groups?**

This question will be addressed by creating an annotation framework using a thematic analysis approach to annotate the WhatsApp dataset according to the conversational functions used. This will be followed by a quantitative and qualitative analysis of conversational function usage patterns amongst male and female users.

3. To what extent can the gender of Kuwaiti social media users be predicted from their online language use?

This question will be explored as a starting point in the WhatsApp study when developing the baseline gender classification system. It will then be explored through a large-scale Twitter study. Tweets in KA will be collected and then a gender classification system for KA will be built using (1) a supervised machine learning approach informed by linguistic features observed in the WhatsApp study and other statistical features (2) a deep learning approach based on fine-tuning a pre-trained large language model using BERT.

3.2 Research Methodology

This section discusses the methodology followed in the three studies conducted that address the research questions. We describe in detail the steps followed in all studies and link the studies with the research question they target.

3.3 Collection and Analysis of features from a WhatsApp Dataset

This study addresses research question 1 in identifying linguistic characteristics of Kuwaiti female and male social media users, and research question 3 in studying the extent to which gender can be predicted from online language usage. The methodology that has been followed in this study is described below:

- **Data Collection:** Ethical approval was obtained before conducting the study. Potential participants from WhatsApp reading club groups were contacted through WhatsApp. The aims and objectives of the study were explained to the potential participants. They were notified that their data will be protected by anonymising their identities. They were given a chance to ask any questions. WhatsApp members who expressed interest in participating in the experiment were sent an information sheet that contains all necessary information regarding the study and a consent form to read and confirm

their interest in participating. The researcher was then added to the WhatsApp reading club groups and exported the chats after 9 months of being added. The data was exported from the researcher's mobile phone and stored in the COM departmental servers (anonymised form). The data was deleted from the researcher's mobile phone directly after exporting it.

- **Data Pre-processing:** All sensitive and personal information were removed, URL links, digits, diacritics and punctuation were removed. Real names mentioned in the chats were replaced with fictitious names. Tokenisation was implemented.
- **Data Analysis** A qualitative and quantitative analysis was carried out to analyse interactional features such as average utterance length, number of turns, average emoji usage, and gender exclusive words.
- **Feature Extraction:** Features inferred from the qualitative and quantitative analysis were extracted and tested in a baseline gender classification system. Further details are presented in **Sec. 4.5.2**.
- **Training:** The features were fed into the supervised machine learning classifiers with their corresponding gender labels. Different machine learning classifiers were tested such as Support Vector Machines, K Nearest neighbour, Logistic Regression and Decision Trees.
- **Testing:** The trained models were tested using evaluation metrics commonly used in classification tasks such as accuracy, and balanced-accuracy. We used balanced-accuracy due to having imbalanced data.

3.4 WhatsApp Dataset Annotation for Conversational Analysis

This study addresses research question 2 in identifying conversational strategies employed by Kuwaiti male and female users in the WhatsApp exchanges. It builds upon the dataset compiled in the study presented in **Sec. 3.3**. The methodology that has been followed is explained below:

- **Annotation Framework:** In order to study the conversational functions that have been used in the WhatsApp dataset compiled, an annotation framework was developed. This involved conducting an inductive thematic analysis (Braun and Clarke, 2006) to

infer the themes/conversational functions employed. The taglist was built using the themes derived and definitions for all conversational function tags were given. A set of annotation guidelines was created for annotators to follow when annotating the dataset.

- **Dataset Annotation:** Annotators who are native speakers of Kuwaiti Arabic were recruited and trained to annotate the dataset using an online automatic annotation tool. They were asked to follow the guidelines to tag each sentence with its corresponding conversational function. Inter-annotator agreement scores were computed and a final consensus annotation was created.
- **Dataset Analysis:** The annotated dataset was then analysed quantitatively using statistical tests such as the Mann-Whitney and Chi-Square tests and qualitatively by describing significant qualitative observations.

3.5 Gender Classification Using KA Tweets

This study addresses research questions 1 and 3. It involves compiling a dataset of Kuwaiti Arabic tweets labelled by gender and using this dataset to build a gender classification system. Question 1 is addressed by analysing linguistic differences between men and women in the dataset. Question 3 is addressed by building a gender classification system to study the extent to which we can predict the gender of social media users from their online language usage. We approach the gender classification task using two supervised learning approaches: a feature engineering approach and a deep learning approach based on fine-tuning a pretrained LLM. The methodologies followed in both approaches are described below:

3.5.1 Feature Engineering Approach

In this approach, features are pre-determined by the researcher and used in the feature extraction stage. We adopt this approach as we hypothesise it will give us insights into what linguistic features in the KA dialect used in Twitter are gender indicative. Using this approach also allows us to test the features inferred in our previous studies and assess how well they perform in predicting the gender leading to a better understanding of the KA dialect in relation to gender.

- **Data Collection:** Ethical approval was obtained before conducting the study. An advertisement was created that sought Kuwaiti female and male Twitter users who tweet in Kuwaiti Arabic. This request specified users who actively tweet in Kuwaiti Arabic and were willing to grant consent for the collection and publication of their

tweets. Upon receiving consent from participants, the researcher used the Twitter academic API (Application Programming Interface) to retrieve the tweets of the consenting users. Gender labels were assigned to the collected tweets based on the information provided by the users during their consent process.

- **Data Pre-processing:** The compiled dataset underwent pre-processing. Tweets were anonymised by replacing the usernames with @USER. Tweets were then tokenised, stop words were removed, hashtags and mentions were filtered out, digits, links, punctuation and diacritics were removed and words were normalised using different normalising techniques used when preprocessing the Arabic language.
- **Feature Extraction:** Different linguistic and statistical features were extracted from the tweets and fed into supervised machine learning classifiers. Further details about features extracted are presented in **Sec. 6.5.3**.
- **Training:** The features extracted were used to train four different supervised machine learning classifiers: Support Vector Machines, K Nearest neighbour, Logistic Regression and Decision Trees.
- **Testing:** The performance of the supervised machine learning classifiers was evaluated using the accuracy metric as the dataset used is a balanced dataset.

3.5.2 Deep Learning Approach

Since the state-of-the-art approaches to classification tasks is using a deep learning approach and more specifically, using Transformers, we wanted to try this approach and assess how well it works in predicting the gender of Kuwaiti users of Twitter. The results obtained from the deep learning model was compared with those derived from the feature engineering approach. This comparative analysis aims to provide insights into the efficacy of each method. Specifically, it will help us understand whether our manually engineered features capture the differences in language of Kuwaiti men and women as effectively as the deep learning model, or if the deep learning model reveals features that were not explicitly used in our manually crafted approach.

In our deep learning approach, three pre-trained language models from Hugging Face (Wolf et al., 2019): an open source repository of large-scale pre-trained transformer models, were fine-tuned using the Ktrain library (Maiya, 2022). We fine tuned CAMELBERT-DA (Inoue et al., 2021b), Marbert (Abdul-Mageed et al., 2021) and QARiB (Abdelali et al., 2021) to be used in our gender classification task. We then evaluated the performance of the model using accuracy scores as evaluating metrics.

Chapter 4

Collection and Computational Analysis of Linguistic Differences Amongst Men and Women in a Kuwaiti Arabic WhatsApp Dataset

4.1 Introduction

This study focuses on the collection and computational analysis of Kuwaiti Arabic, which is considered a low resource dialect, to test different sociolinguistic hypotheses related to gendered language use. In this study, we describe the collection and analysis of a corpus of WhatsApp Group chats with mixed gender Kuwaiti participants. This corpus, which we make publicly available, is the first corpus of Kuwaiti Arabic conversational data. We analyse different interactional and linguistic features to get insights about features that may be indicative of gender to inform the development of a gender classification system for Kuwaiti Arabic in the study presented in Chapter 6.

This study contributes to the field of ANLP in two ways. First, we have compiled and made publicly available a new, gender-labelled KA dataset (KAGen), which can be used by researchers interested in the Kuwaiti dialect or gender studies. This dataset consists of textual book club conversations conducted on the WhatsApp online instant messaging mobile application. To the best of our knowledge this is the first published dataset of mixed gender KA conversational data. Second, we have carried out an analysis of interactional and linguistic features that is used to inform the development of a gender classification system for KA in Chapter 6.

In Sec. 4.2, we review the related work. In Sec. 4.3 we describe the methodology followed in conducting this study. In Sec. 4.4, we analyse the results qualitatively and quantitatively. In Sec. 4.5, we test our features on a baseline gender classification system and evaluate its performance. Finally, in Sec. 4.6, we conclude and summarise the work done in this study.

4.2 Related Work

In the context of studies that have explored gender differences in language use, Rosenfeld et al. (2016) looked into gender differences in language usage of WhatsApp groups. They analysed over 4 million WhatsApp messages from more than 100 users to find and understand differences between different age and gender demographic groups. In analysing the data, they relied on metadata only such as message lengths, size of the WhatsApp groups, time, average number of sentences sent per day, time between messages. In relation to gender, analysing the length of messages sent by both genders showed that women tend to send longer messages than men. On average, women's messages contain 6.5 words, while men's messages contain 5.2 words. They also concluded that women are more active in small WhatsApp groups, whereas men are more active in larger WhatsApp groups. These differences were then employed in building age and gender prediction models. They performed a 10-fold cross validation for these tasks using decision trees and a Bayesian network. For the gender prediction task, using users' metadata with decision trees achieved 0.70 accuracy and 0.73 accuracy when used with a Bayesian network. Moreover, Ghilzai and Baloch (2015) explored gender differences in turn-taking across various conversational contexts, such as radio programs, TV shows, and casual conversations. The research found that while there is no significant difference in turn-taking rates within same-gender conversations (male-to-male or female-to-female), women tend to take more turns in mixed-gender interactions.

Furthermore, in the context of email messages, Thomson et al. (2001) conducted experiments to analyse how men and women interact linguistically. The first experiment involved 11 male and 11 female participants all of whom were assigned two netpals (1 female username and 1 male username) who were actually one person that used a female-preferential language when using a female username and a male-preferential language when using a male username. Thirteen different linguistic features from the messages were analysed such as references to emotions, opinions, apologies, adjectives and other features. The mean frequencies of these features were calculated. Results showed that the language style used by the netpal, not the gender labels of the participant, influenced the language used by the participant. This means that participants would accommodate their language to the language of their corresponding

netpals in their responses, regardless of their own gendered language style. In the second experiment, 33 female participants and 32 male participants were paired with one netpal. The netpals communicated in 4 ways: 2 in which the username labels matched the language style of the gender label and 2 in which the username labels do not match the language style of that gender (e.g., Kate is the label and the language used was male preferential language). Results showed that there was an effect of the participants' gender in the language they used in cases where the labels did not match the language style. "Given that convergent accommodation is a sign of liking and acceptance, participants might have been signaling non-acceptance by maintaining their own gender-preferential style when netpals' style and gender did not match" (Thomson et al., 2001, p.174).

Other studies have looked into differences amongst genders in the use of emojis. Chen et al. (2017) compiled a large dataset of 401 million smartphone messages in 58 different languages and labelled them according to the gender of users. They used emojis from the dataset to study how they are used by males and females in terms of emoji frequency, emoji preference and sentiment conveyed by the emojis. They also studied the extent to which emojis are indicative of gender when used in a gender classification system. The results obtained from this study showed that not only are there considerable differences in the use of emojis between males and females, but also that a gender classification system that uses emojis alone as features can achieve an accuracy of 0.81.

Shared NLP tasks that are organised for the research community have started off by tackling problems with the English language and in recent years have added Arabic datasets, reflecting the increasing interest in Arabic NLP. For example, the PAN 2017 Author Profiling Shared Task included two tasks: gender identification and language variety identification of Twitter users. Arabic, English, Portuguese, and Spanish datasets consisting of tweets were provided for training and testing. The system that achieved the highest accuracy result on gender identification in the Arabic dataset was the system developed by Basile et al. (2017). They used an SVM classifier in combination with word unigrams and character 3- to 5-grams and achieved an accuracy of 0.80.

As for studies that have targeted the Arabic language, Alsmearat et al. (2014) studied gender text classification of Arabic articles using the Bag-of-Words (BoW) approach. They collected and manually labelled 500 Arabic articles from different Arabic news websites. The number of articles was distributed equally across both genders. They wanted to explore the result of performing feature reduction techniques such as PCA and correlation analysis on the high-dimensional data in combination with different machine learning algorithms for the gender classification task. Results showed that Stochastic Gradient Descent (SGD),

Naive Bayes Multinomial (NBM) and Support Vector Machines (SVM) were the classifiers that performed best on the original dataset where the accuracy results surpassed 0.90.

Furthermore, Mubarak et al. (2022) compiled a dataset of 166K Arabic tweets and labelled them with gender and geo location labels. They used this dataset for gender analysis and to build a gender classification system using SVMs that was tested on different features such as usernames of the twitter users, the profile pictures of the users, tweets and gender distribution of users' friends. Their study showed that using usernames alone as features for gender prediction achieved the highest F1 score of 0.82. In addition, Hussein et al. (2019) attempted to build a gender classification system for Egyptian Arabic. They created a dataset of 140K tweets that were retrieved from famous Egyptian influencers and active Egyptian users of Twitter. They labelled the dataset according to the gender of the Twitter users by referring to the users' profile image and names. They experimented with different features such as gender discriminative emojis, female suffixes, manually created dictionaries of swear words, emotion words, political words, flirting words, technological words and word embeddings. They used ensemble weighted average on a mixed feature vector fed into a Random Forest classifier and an N-gram feature vector fed into a Logistic Regression classifier. They achieved an accuracy score of 0.88.

In addition, Arafat and Hamamra (2021) investigated the use of word elongation in Facebook-mediated communication in Palestinian Arabic to explore how men and women use word elongation to convey emotions and social identity. The study revealed that women employ word elongation more frequently to convey positive emotions, while men use it less often, and primarily to express anger when they do. Not many gender studies in NLP have provided much insight into linguistic characteristics of gendered language, especially those related to dialectal Arabic. Furthermore, the field of ANLP still lacks enough dialectal arabic datasets to help inform the development of Arabic natural language processing tools. We conduct a pilot study in this chapter to collect conversational KA and analyse interactional features to inform the building of a gender classification system for KA.

4.3 Methodology

4.3.1 Data Collection

Since we are interested in studying the features of conversational data of Kuwaiti men and women, we chose to collect textual data from WhatsApp reading club groups. To ensure the publication of the conversations while adhering to the ethical considerations, we determined that the most appropriate platform would be WhatsApp reading club groups as they feature

written Kuwaiti Arabic and focus on discussions that do not involve sensitive topics or personal information.

As part of the data collection process, we applied for ethical approval before conducting the study. This involved ensuring that all participants were aware of the nature and purpose of the study and their role in it. We obtained informed consent from all participants.

The dataset was collected from three Kuwaiti reading club WhatsApp groups. These were already existing WhatsApp reading club groups that have been running for years and are managed by Kuwaiti admins. All participants were native Kuwaiti speakers whose first language is KA. The researcher was added to the groups to be able to export the chat after 9 months of being added. The chats were then exported from the mobile phone and saved in the researcher's computer for processing.

The dataset which we name KAGen, consists of 4479 turns (2623 turns by females and 1856 turns by males), see Table. 4.1. The dataset is made publicly available for researchers in the research field.

Gender	Number of Turns	Word Count
Women (28 participants)	2623	17388
Men (14 participants)	1856	14005
Total	4479	31393

Table 4.1 Descriptive Statistics of the KAGen Dataset

4.3.2 Data Preprocessing

A number of steps were taken prior to exporting the chats from the researcher's mobile. This involved anonymising the names of the WhatsApp members. The usernames were replaced with the word "USER" concatenated with a number and a letter to represent the gender of the user (e.g, USER1F). The chats were then exported to the researcher's computer to prepare the data for computational processing. The following preprocessing steps were performed:

1. All sensitive and personal information was removed.
2. Real names that were mentioned in the chat were replaced with fictitious names.
3. URL links were removed.
4. Two versions of the dataset were created using the CAMel tools, built by Obeid et al. (2020), for preprocessing: one that involves tokenisation, removal of digits,

diacritics and punctuation and changing alef variants to **l** and alef maksura to **ي** and teh marbuta to **o**; and another version that involves tokenisation and punctuation removal. Depending on the type of textual analysis required, the dataset version was chosen.

4.3.3 Feature Analysis

We were interested in exploring interactional features and lexical features pertaining to the KA dialect. We chose to study how the following features were used amongst men and women participating in the study:

- Length of turns per gender (word count):
Inspired by Rosenfeld et al. (2016) who found that women tend to send longer messages than men, with a significant difference in the average word count per message. We aimed to determine if this pattern holds in the KA dialect.
- Number of turns per gender:
Building upon the analysis of length of turns, we wanted to investigate the number of turns amongst men and women in the collected WhatsApp groups to examine the level of activity amongst men and women.
- Use of emojis amongst females and males:
Inspired by Chen et al. (2017), who found significant gender differences in emoji frequency, preference, and sentiment across various languages and effectiveness if used a feature for gender classification.
- Whether there are KA words or expressions that are exclusive to each gender:
This stems from the researcher's observations in real-life communication within the Kuwaiti society.
- Most frequently used words:
As some studies looked into the most frequently used words by men and women such as Rayson et al. (1997) and Koppel et al. (2002), we wanted to analyse the most frequently used words in our dataset.
- Lengthened or elongated words:
We chose to explore the use of elongated words due to their significance in conveying emotions and interest to examine whether our findings are similar or different from (Arafat and Hamamra, 2021).

Table 4.2. presents the descriptive statistics of the first three features.

Gender		Emoji Count	Word Count	Num of Turns
Women (28 participants)	Total Number	2144	17388	2623
	Mean	76	621	94
	Median	23	163	29
	Std. Deviation	123	1132	144
	Minimum	2	6	2
	Maximum	506	5611	655
Men (14 participants)	Total Number	801	14005	1856
	Mean	57	1000	133
	Median	36	432	102
	Std. Deviation	68	1197	134
	Minimum	1	5	3
	Maximum	249	3941	444

Table 4.2 Descriptive Statistics of the Features Analysed

4.4 Quantitative and Qualitative Analysis of Chats

To analyse the results of this study, two approaches were taken: a quantitative statistical approach and a qualitative linguistic approach. The statistical analysis was done using SPSS¹. One limitation of using a statistical approach in analysing the data is that it does not take into account the contextual information and meanings embedded within the text. Therefore, it was important to perform an in-depth manual analysis of the data to be able to describe the patterns found and provide interpretations for points that the statistical analysis could not capture.

4.4.1 Quantitative Analysis

Our quantitative analysis aims to explore potential differences between men and women across three key features: number of turns, average utterance length, and average emoji usage. Initially, we encountered challenges stemming from data distribution inconsistencies and the presence of outliers within our features.

Due to having more females than males, we opted to perform normalisation rather than analysing raw counts to have more accurate interpretations. We normalised the feature values of words counts and emoji counts by dividing them by the total number of turns for each user and then applied a square root transformation. This approach scales the features relative

¹Statistical Package for the Social Sciences: a statistical analysis software package. <https://www.ibm.com/products/spss-statistics>

to the number of turns, ensuring that each user's communicative behavior is evaluated in proportion to their level of engagement.

Due to the complexity and variability of human language use, language data usually shows non-normalities such as skewed distributions, high variance and presence of outliers. We chose to retain most outliers because they could potentially reflect their natural textual communication patterns. However, we made an exception for WhatsApp group admins. Admins have specific responsibilities such as answering questions, sending reminders, and enforcing rules, which may lead to an atypical amount of text production not reflective of their natural communicative behavior. As all admins are females, we replaced their feature values with the mean of the female group if they were identified as outliers.

In our analysis below, we initially employ a box plot to assess the distribution of the data and identify any outliers. We then examine whether the outliers are admins or not. If the outliers are admins, we address them by substituting their values with the mean of their respective gender group.

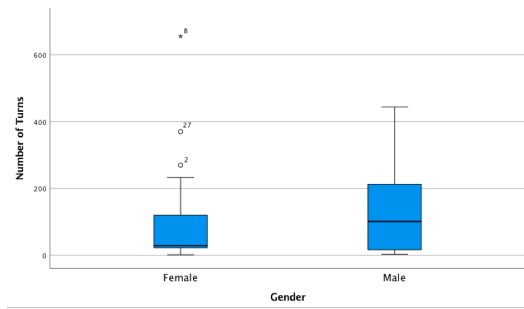
Following outlier treatment, we proceed to conduct a normality test to evaluate whether the data conforms to a normal distribution. If the data is normally distributed, we conduct a t-test for our statistical comparison. However, if the data is not normally distributed, we opt for the Mann-Whitney U test (Mann and Whitney, 1947).

In our significance testing, we interpret a calculated p-value less than 0.05 as strong evidence of a difference between men and women. If the p-value is less than 0.1, we consider it indicative of a difference, albeit less strong than the conventional threshold. In cases where the calculated p-value falls between the two thresholds (0.05 and 0.1), we report findings for both the 0.05 and the 0.1 significance levels.

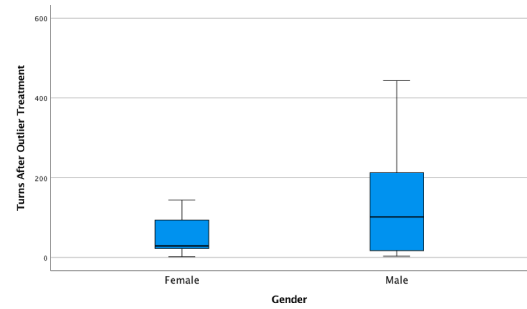
Number of Turns

To begin our analysis, we focused on examining the number of turns, which would serve as the denominator for normalising word counts and emoji counts. However, before proceeding, it was crucial to address any outliers among the group admins that appeared in the distribution of the number of turns.

As can be seen in Fig. 4.1a, which displays the distribution of total number of turns amongst males and females, outliers (users 2, 8 ,and 27) were identified, all of whom were female admins of the WhatsApp groups. These outliers were subsequently treated by replacing their values with the female mean of (94). Following this treatment, a second box plot was generated to visualise the distribution of total number of turns after outlier treatment. See Fig. 4.1b in which there are no outliers.



(a) Distribution of Males and Females Total Number of Turn Counts



(b) Distribution of Males and Females Total Number of Turn Counts (Outliers Treated)

Figure 4.1 Comparison of total number of turns before and after treating outliers

We performed square root transformation for the number of turns after treating the outliers and plotted their distribution in a box plot as shown below in Fig. 4.2.

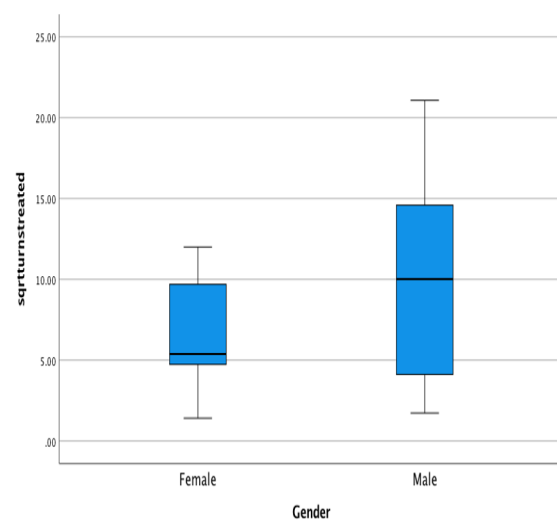


Figure 4.2 Distribution of Square Root Transformation Applied to Number of Turns

We then conducted the Shapiro-Wilk normality test to check the distribution of the square root of number of turns feature after normalisation, taking into account our sample size (less than 50). The p-values obtained from the Shapiro-Wilk test for the square root transformed turn counts in both gender groups are 0.123 and 0.667, respectively. For the Shapiro-Wilk test, the null-hypothesis is that the population is normally distributed. Therefore, these values being higher than the conventional significance level of 0.05 suggest the null hypothesis cannot be rejected and that therefore the data may well be normally distributed. Given this observation, we use the t-test in our statistical analysis.

Test of Normality			
Gender	Shapiro-Wilk		
	Statistic	df	Sig.
Female	.935	28	.123
Male	.957	14	.667

Table 4.3 Test of Normality for Square Root Transformed Turn Counts

As the number of turns could be indicative of the level of interaction in the WhatsApp groups, we formulate the following hypotheses:

the null hypothesis (H_0):

H_0 : there is no difference between men and women in their level of interaction measured through their number of turns.

and an alternative hypothesis (H_1):

H_1 : there is a difference between men and women in their level of interaction measured through their number of turns.

Gender	N	Mean	SD
Female	28	6.34	2.91
Male	14	9.92	6.05
T-test P-value		0.013	

Table 4.4 T-Test Results for Square Root Transformed Turn Counts

Based on the conducted t-test, the results in Table 4.4 indicate that there is a statistically significant difference in the level of interaction, as measured by the number of turns, between men and women in the WhatsApp groups. The null hypothesis (H_0) suggests that there is no difference between men and women in their level of interaction. However, with a p-value of 0.013, which is below the conventional significance level of 0.05, we reject the null hypothesis. This rejection supports the alternative hypothesis (H_1), which states that there is indeed a difference between men and women in their level of interaction.

Specifically, the mean number of turns for females is 6.34, with a standard deviation of 2.91, while for males, the mean is 9.92, with a standard deviation of 6.05. The lower mean and standard deviation for females compared to males suggest that females, on average, engage in fewer turns compared to males, indicating potentially differing levels of interaction within the WhatsApp groups.

Average Utterance Length

The length of turns, measured by the total word counts, was computed to examine if men and women differ in this feature. We calculated the total number of words for each user in our dataset, as detailed in Table 4.2. To normalise the word count feature, for each participant we took the square root of dividing the participant's raw word count by the number of their turns, i.e. the square root of their mean word count per turn. Following the treatment of outliers in the turns distribution, Fig. 4.3 displays the box plot of the square root transformation applied to word counts divided by the number of turns. Although two outliers were detected, they were not treated as they did not belong to administrator accounts.

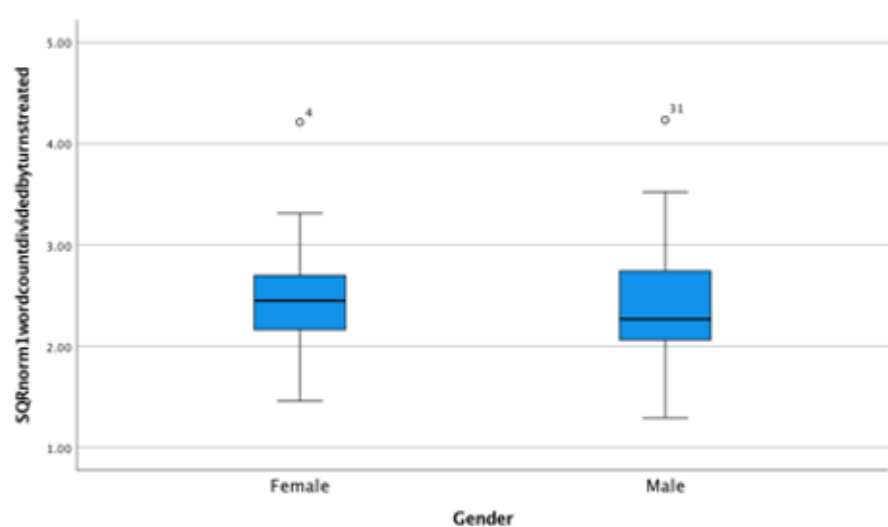


Figure 4.3 Distribution of Square Root Transformation Applied to Word Counts Divided by Number of Turns

Test of Normality			
Gender	Shapiro-Wilk		
	Statistic	df	Sig.
Female	.946	28	.158
Male	.888	14	.042

Table 4.5 Test of Normality for Square Root of Word Counts Divided by Number of Turns

The Shapiro-Wilk test results, as shown in Table 4.5, reveal a significance level of 0.158 for the female group and 0.042 for the male group. This suggests non-normality in the distribution of the square root of word counts divided by the number of turns for the male

group, while the distribution for the female group appears to be approximately normal. Given the non-normality in the male group and the differing distributions between genders, we used the Mann-Whitney U test as it does not rely on the assumption of normality.

The null hypothesis (H_0):

H_0 : there is no difference between men and women in their average utterance length measured through their number of words per turn.

and an alternative hypothesis (H_1):

H_1 : there is a difference between men and women in their average utterance length measured through their number of words per turn.

Gender	N	Mean Rank	Sum of Ranks
Female	28	22.36	626.00
Male	14	19.79	277.00
p-value	0.522		

Table 4.6 Mann-Whitney Statistics for Square Root of Word Counts Divided by Turns

Based on the results obtained using the Mann-Whitney U test (p-value = 0.522, U = 172.00) for average utterance length, we find that the p-value is greater than the significance level of 0.05. Therefore, there is insufficient evidence to conclude that there is a significant difference in the average utterance length between men and women.

Average Emoji Usage

We also analyse how likely it is for men and women to use emojis when interacting in the WhatsApp chat groups. We noticed that on average women used .82 emojis per turn, while men used on average .43 emojis per turn. Therefore, the odds of using emojis amongst women compared to men is 1.9:1, indicating that women were almost 2 times more likely to use emojis than men. Further analysis involved normalising the number of emojis feature by dividing it by the number of turns. We conducted a square root transformation and then plotted the distribution of the normalised emoji counts.

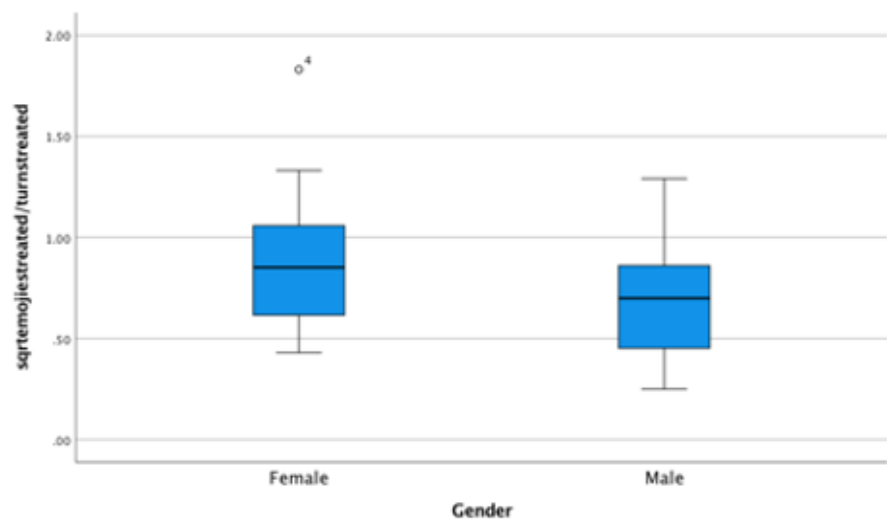


Figure 4.4 Distribution of Square Root Transformation Applied to Emoji Counts Divided by Number of Turns

As can be seen in Fig. 4.4, one outlier was detected in the female group. However, the outlier is not an admin and therefore we did not treat her. We then conducted the Shapiro-Wilk normality test.

Test of Normality			
Gender	Shapiro-Wilk		
	Statistic	df	Sig.
Female	.938	28	.100
Male	.961	14	.734

Table 4.7 Test of Normality for Square Root of Emoji Counts Divided by Number of Turns

Given that the p-values in both groups exceed the threshold of 0.05, indicating approximate normal distribution, the t-test is selected to compare the means of the two groups. We formulate the following hypotheses:

the null hypothesis (H_0):

H_0 : there is no difference between men and women in their average emoji usage.

and an alternative hypothesis (H_1):

H_1 : there is a difference between men and women in their average emoji usage.

Gender	N	Mean	SD
Female	28	0.869	0.322
Male	14	0.684	0.276
T-test p-value		0.074	

Table 4.8 T-Test Results for Square Root Transformed Emoji Counts

The T-test results for the square root-transformed emoji counts, as shown in Table 4.8, indicate interesting insights into the emoji usage patterns between men and women. With a significance level threshold of 0.05, the two-sided p-value significance of 0.074 for both men and women suggests a lack of statistical significance between them in their use of emojis. However, with a significance level threshold of 0.10, the two-sided p-value significance of 0.074 for both men and women indicates a statistical significance between them in their use of emojis, with women showing a higher usage compared to men. This implies that there is some evidence that women tend to use more emojis than men.

Summary of Results

The analysis of the three features: number of turns, average utterance length, and average emoji usage within the WhatsApp groups revealed interesting gender-based differences. Specifically, that men may contribute more frequently to the conversations in the WhatsApp groups as seen in the distribution of their number of turns. This finding contrasts Ghilzai and Baloch (2015) study which found that women tend to take more turns in mixed-gender conversations across formal and casual settings. In terms of emoji usage, women displayed a higher usage compared to men, suggesting that women are more expressive through emojis and may use them as a means of enhancing communication or conveying emotions within the groups. This finding contrasts with Chen et al. (2017), who observed that while women used more emojis on public platforms like Twitter, men tended to surpass women in emoji usage in private communication settings such as WhatsApp. However, when examining average utterance length, the analysis did not yield sufficient evidence to suggest a significant difference between men and women, which contrasts with Rosenfeld et al. (2016), where their findings showed that women tend to send longer messages than men.

4.4.2 Qualitative Analysis

We were interested in qualitatively analysing some features experimented with. First, we wanted to explore the types of emojis, exclusivity of emojis and patterns of emojis used by

men and women to achieve a better understanding of emoji usage amongst genders. This is discussed in the following section.







































Rank	Women		Men	
	Emoji	Count	Emoji	Count
1		218		156
2		211		143
3		193		43
4		140		42
5		116		28
6		95		26
7		91		20
8		85		18
9		75		17
10		75		15

Table 4.9 Top Ten Emojis Used by Kuwaiti Women and Men

Frequency and Types of Emojis

Emojis were significant features observed in the group chats and were commonly used by both men and women. Women used a total of 2144 emojis, while men used a total of 801 emojis. As for the types of emojis used, various differences were observed. Emojis used by women are from a wide range of emoji categories and are colorful, whereas men used a limited set of emojis from certain categories. 68% of women used heart emojis, whereas only 29% of men used heart emojis. It was also noticed that women used different types and colors of heart emojis. However, men used limited heart emojis , , . Further more, women used a large variety of flowers and plants , , , , , , , , whereas men used only two types of flowers  and .

The analysis also involved computing the 10 most frequently used emojis by men and women as shown in Table 4.9. As it can be seen, the top used emojis for both men and women are ( and ) which shows that both men and women are encouraging and applauding each other. It was observed that men used ( and ) significantly more than all the other emojis extracted, which were mainly smileys. In comparing the top 10 lists of emojis by men and women, it was noticed that women used  (193 times) notably higher than men (15 times) and used flowers more than smileys as opposed to men.

Exclusivity of Emojis

There are some stereotypes regarding emoji usage such as that there are certain emojis that are not used by men due to them implying a feminine sense and other emojis not used by women because they are masculine. This study examined this stereotype to explore if this can be considered a feature indicative of gender. The emojis that were exclusively used by each gender were extracted and compared. It was noticed that men refrained from using certain emojis that are stereotypically considered feminine and were used by women in the group chats such as 💋, 😘, 😺, 😊, 💖, 💖, 💖, 💖, 💖, 🦋. This observation also supports the hypothesis that women are more emotionally expressive than men (Goldshmidt and Weller, 2000). The emojis that were exclusively used by men mainly consisted of male character emojis such as 🧑, 🧑, 🧑, 🧑, 🧑.

Patterns of Emoji Usage

A number of observations were made related to patterns of emoji usage. Women used a larger variety of emojis across different categories (smileys and people, activity, travel and places, food and drink, nature .. etc) than men to express themselves. Men used limited types of emojis from certain categories (smileys and people, nature) and very limited use of hearts or emojis that express emotions.

A pattern was also noticed regarding the number of emojis used per turn. Most users used one or two emojis in a turn and this lead to interest in analysing bigrams of emojis used by men and women to explore if there are any patterns of use or certain emoji combinations used. The most frequently used bigrams consisted of the same emoji repeated rather than a combination of two different emojis. It was observed that certain combinations were used significantly more by each gender. For example, 😂😂 was used 70 times by men and 38 times by women, 😍😍 was used 3 times by men and 64 times by women, and 🙌🙌 was used 4 times by men and 80 times by women. This showed certain emoji combinations may be used with different frequencies amongst men and women.

KA Lexical choices and Features

Other exploratory data analysis was conducted to analyse the lexical choices amongst men and women in the WhatsApp groups. Features such as the most frequently used words, the exclusively used words and other lexical features were analysed.

Analysis regarding the most frequently used words showed that the word “Allah”, الله was one of the highly repeated words amongst both men (262 times) and women (325 times). “Allah” means “God” and could appear in a sentence as a separate word or part of a phrase

such as “masha’Allah”, ماشاءالله which is an expression used to express appreciation when someone hears good news, and “inshaAllah”, ان شاء الله which is an expression used to convey willingness to do something. The high repetition of these phrases could indicate cooperativeness and politeness in the conversations. The word “alketab” الكتاب which means “book” was also amongst the highest repeated words amongst men (32 times) and women (99 times). This is due to the conversations mainly revolving around reading books. Figure 4.5 and Figure 4.6 show the most frequent words in both the women’s and men’s chats.

Analysis was also done on the exclusively used words amongst both men and women. One aim of extracting the gender exclusive words was to find KA gendered words that denote femininity or masculinity to inform the development of a gender classification system. However, due to the formal nature of the WhatsApp reading club groups, only a few examples of this phenomenon were captured and they were mostly in women’s messages. Some of the examples of female exclusive words found are: “shatoora” شطورة, meaning “smart girl”, “b’khatri” بخاطري meaning “I really want ..”, “habeebty” حبيبتي, meaning “my dear”, “s’ghairoona” صغيروونه, meaning “very small”, “katkoota” كتكووته, meaning “so cute” and “please” بليز.

Analysis of the chat also showed high occurrence of lengthened or elongated words which are words that include repeated letters to emphasise different meanings such as هههههههه “hhhhhhhhh” expressing laughter and وaaaaا “woooooow” expressing amazement. Lengthened words can be indicators of expressing feelings which is stereotypically attached to women’s speech, and therefore we wanted to test this hypothesis by determining the number of lengthened words used by men and women per turn on average. There were some interesting observations. Women used 0.057 lengthened words per turn on average (so about once per 18 turns), whereas men used 0.037 (about once in 28 turns). This indicates that women tend to lengthen words roughly 1.5 times as often as men. After performing further inspection to the lengthened words, it was observed that women tend to perform this with a large variety of words when laughing هههههههه “hhhhhh”, complimenting حمممييه “beautifuuul”, congratulating مبرووووك “congraaatulations”, encouraging براااقووووو “bravooooo”, agreeing ايي “yees”, greeting صباح النووور “good mooorning” and expressing feelings such as missing the members مشتاقيين “miiiis you”. However, men’s use of lengthened words were less diverse. They mostly used lengthening when laughing ههههههههههههههه “hhhhhhhhhhhhhhhh” and greeting هلاا “hiii”.



Figure 4.5 Most Frequent Words Used by Women



Figure 4.6 Most Frequent Words Used by Men

4.5 A Baseline Gender Classification System

Following the work done in collecting and analysing the KA WhatsApp dataset, we wanted to test how effective these features are in automatically identifying the gender of the users. To experiment this, we fed the features to a baseline gender classification system using a supervised machine learning approach. We used stratified 10-fold cross-validation to ensure a representative distribution of classes in both the training and testing sets. Four common supervised machine learning classifiers used for text classification tasks were used: K-Nearest Neighbours Algorithm (KNN), Support Vector Machines (SVM), Logistic Regression (LR),

and Decision Trees (DTs). The dataset which consists of 4479 turns (2623 turns by women and 1856 turns by men) was preprocessed. Details of the preprocessing techniques that were applied are listed in the following subsection.

4.5.1 Data Preprocessing

We performed sentence splitting to the dataset. This resulted with 5289 sentences (3121 sentences by women and 2168 sentences by men). To test the effect of preprocessing on the performance of the classifiers, the following techniques were applied:

1. the sentences were tokenised.
2. punctuation was removed.
3. hashtags, @, and digits were removed.
4. URL links were removed.
5. diacritics were removed.
6. elongated words were converted to their normal forms.
7. alef variants were converted to ا.
8. alef maksura was converted to اِي.
9. teh marbuta was converted to ه.

4.5.2 Feature Extraction

In this baseline gender classification system, we extracted the features discussed in our study using four different supervised learning classifiers. We extracted word counts, emoji bigrams, sentence lengths, and stretched words. The performance of each feature was evaluated using 10-fold cross-validation. Specifically, we assessed the classifiers' accuracy (Acc) and balanced accuracy (BAcc) when each feature is used separately or in different combinations. Description of each feature extracted and classification results are presented below.

Bag-of-Words (BOW)

To extract word counts as a feature we used a BOW approach. we used `CountVectorizer` from Scikit-learn to create a matrix of word counts for the sentences in our dataset. The `fit_transform()` method from Scikit-learn was used to learn the vocabulary (unique words) present in the entire dataset and extract individual term-frequency vectors for each sentence, representing the count of each word in the learned vocabulary. The resulting list of term-frequency vectors is then fed into the machine learning classifiers.

Emoji Bigrams

As patterns of emoji use were observed to be different amongst men and women in the qualitative analysis of the study, we wanted to test emoji bigrams as features using the baseline gender classification system developed. We represented this feature following the BOW concept but in this case we were counting the occurrence of emoji bigrams and creating a matrix of emoji bigram counts. The resulting emoji bigram matrix was fed into the machine learning classifiers.

Length of Turns

We wanted to use the length of turns as features. This was done by counting the number of words in every turn and saving it in a vector. This resulted with a final feature vector that contains the total word count for each sentence in the dataset. This feature vector is then fed into the machine learning classifiers.

Stretched Words

We wanted to test the linguistic phenomena of stretched words in the gender classification system. We extracted stretched words features in two ways. We considered words with letters repeated from 3 till 10 times stretched. The first representation involved counting the total number of stretched words in every sentence and passing a vector of the total counts for all sentences to the machine learning classifiers. The second representation involved assigning a binary value for every sentence according to whether a stretched word occurs in the sentence or not. 0 is given to sentences that do not include stretched words and 1 is given to sentences that include stretched words.

4.5.3 Results and Analysis

Features	KNN				SVM				LR				DT			
	PP		UP		PP		UP		PP		UP		PP		UP	
	Acc	BAcc	Acc	BAcc	Acc	BAcc	Acc	BAcc	Acc	BAcc	Acc	BAcc	Acc	BAcc	Acc	BAcc
word count	0.62	0.60	0.61	0.56	0.67	0.65	0.68	0.65	0.68	0.65	0.69	0.66	0.63	0.60	0.66	0.63
sent len	0.53	0.50	0.53	0.51	0.58	0.51	0.58	0.50	0.58	0.50	0.58	0.50	0.58	0.51	0.59	0.52
stretched words	0.53	0.50	0.51	0.50	0.58	0.50	0.58	0.50	0.58	0.50	0.58	0.50	0.58	0.50	0.58	0.50
emoji bigrams	0.51	0.52	0.57	0.51	0.59	0.51	0.59	0.51	0.59	0.51	0.59	0.51	0.58	0.50	0.59	0.50
comb.1	0.54	0.51	0.53	0.51	0.59	0.52	0.59	0.51	0.59	0.51	0.59	0.51	0.57	0.51	0.59	0.53
comb.2	0.61	0.60	0.61	0.61	0.68	0.66	0.69	0.66	0.69	0.66	0.70	0.67	0.64	0.62	0.65	0.63

Table 4.10 Accuracy (Acc) and Balanced Accuracy (BAcc) Results for the Baseline Gender Classification System Using 10 Fold Cross-validation. PP represents pre-processed text and UP represents unprocessed text. Comb.1 = sent len + stretched words + emoji bigrams. Comb.2 = word count + sent len + stretched words + emoji bigrams.

Results of the features tested are presented in Table. 4.10. We report the balanced accuracy (BAcc) scores as the KAGen dataset is imbalanced. The balanced accuracy is a commonly used evaluation metric that accounts for class imbalance. It is calculated as the average of the true positive rate (TPR) for the positive class and the true negative rate (TNR) for the negative class (Brodersen et al., 2010).

$$\text{BAcc} = \frac{1}{2} \left(\frac{TP}{P} + \frac{TN}{N} \right)$$

where:

$$P = TP + FN, \quad N = TN + FP$$

TP (True Positives) is the number of correctly predicted positive samples, TN (True Negatives) is the number of correctly predicted negative samples, $P = TP + FN$ represents the total number of actual positive samples with FN (False Negatives) being those positive samples incorrectly predicted as negative, and $N = TN + FP$ represents the total number of actual negative samples with FP (False Positives) being those negative samples incorrectly predicted as positive.

As can be seen, the balanced accuracy scores range from 0.50 to 0.67 across various models. Comb.2 (word count + sent len + stretched words + emoji bigrams) achieved the highest scores (accuracy: 0.70 and balanced accuracy: 0.67). The feature that seemed to have the most positive effect in improving the performance of the classifier was the word count feature as when tested separately it achieved 0.66 balanced accuracy. It was also noticed that preprocessing the dataset did not improve the performance of the classifiers. In fact, overall,

the results for unprocessed text were higher. Although some qualitative analysis did show patterns of emoji use to be distinctive amongst men and women, we think that the small size of the dataset might have had a negative effect on the classifiers performance. We aim to test this feature and other features in our large scale study in Chapter. 6 to observe if they improve the performance of the gender classification system.

4.6 Summary

We have described in this chapter (KAGen,) the first publicly available dataset of conversational Kuwaiti Arabic that is labelled by gender. We analysed the dataset by looking into interactional and linguistic features that are performed in mixed gender WhatsApp groups. We described the WhatsApp data collection process and analysed features such as number of turns, average utterance length, average emoji usage and Kuwaiti Arabic lexical features. Statistical analysis reveals distinctions between men and women in their interaction patterns. In the number of turns, men showed higher activity levels than women, whereas in average emoji usage, women used emojis more frequently than men. However, there was no significant difference observed in average utterance length. Furthermore, qualitative analysis of other features such as the range and specific types of emojis used, certain lexical choices and the phenomenon of word lengthening revealed considerable differences between women and men's language use.

We have tested the features on a baseline gender classification system for Kuwaiti Arabic trained and tested on (KAGen) and observed that using word counts on text that did not undergo preprocessing performed best in predicting the gender of users. Insights from this study are taken into consideration in a large-scale gender classification study on a KA Twitter dataset in Chapter. 6 where we will be testing them and other features and approaches to study how well they perform in predicting the gender of KA social media users.

Chapter 5

WhatsApp Data Annotation for Conversational Analysis

5.1 Introduction

In the domain of Arabic Natural Language Processing (ANLP), many projects have focused on collecting and annotating MSA datasets. Nonetheless, projects targeting dialectal Arabic are increasing. However most of the dialectal Arabic data are extracted from social media platforms such as Twitter, Facebook, novels, online forums and so on. Conversational dialectal Arabic found in online communities is still an understudied topic. This type of language is unique in its ability to capture real-time, interactive and informal exchanges between interlocutors. Gathering conversational dialectal Arabic from digital platforms creates diverse opportunities to study, for example, theories in discourse analysis, explore sociolinguistic aspects of the language, create lexicons for the dialect and much more. To the best of our knowledge, such datasets in Kuwaiti Arabic have not yet been collected and made publicly available. Due to our interest in studying how conversational analysis takes place amongst Kuwaiti men and women in online platforms, we aimed to use computational tools to annotate conversational Kuwaiti Arabic used in reading club WhatsApp groups with a specific focus on mapping these conversations to their underlying conversational functions. Written Kuwaiti Arabic, which is increasingly prevalent in online communication, particularly on social media platforms such as WhatsApp, forms a promising opportunity to study gender differences in language use. Therefore, in this study we seek to bridge this significant gap by providing a comprehensive framework for annotating and classifying the diverse conversational functions within these WhatsApp interactions, thereby contributing to a deeper understanding of how conversational functions are performed in a mixed sex setting.

Our study will also contribute to the research field by providing an annotated conversational Kuwaiti Arabic dataset that can allow us, and others, to study how conversational interaction takes place in Kuwaiti Arabic in social media and, in particular, whether or not there are observable differences in conversational behaviour based on the gender of the participants and if so what these differences might be.

5.2 Related Work

Annotated corpora serve as the critical foundation that many NLP systems heavily depend on. The importance of annotations in NLP cannot be emphasised enough as they are needed to train and evaluate certain NLP systems and also develop and evaluate linguistic theories (Fort, 2016). Annotating corpora is done either manually or automatically. According to Neves and Ševa (2021, p.146):

Manual annotation is the task of reading a particular pre-selected document and providing additional information in the form of the so-called annotations. Annotations can occur at any level of a linguistic component, i.e. document, paragraph, sentence, phrase, word or character.

When annotating corpora, certain main steps are followed. First, the corpus that will undergo annotation is gathered. Second, the corpus is pre-processed according to the task. Third, the tagset is created and the first version of the guidelines is developed. Fourth, annotators are provided annotation training. Fifth, annotators annotate a sample from the corpus. Sixth, the annotations are compared by producing inter-annotator agreement scores and assessing the level of agreement that is acceptable. If the level of agreement is not reached, the annotation guidelines are redefined. Finally, the corpus is annotated following the final version of the guidelines and a final consensus version of the corpus is created (Fort, 2016; Neves and Ševa, 2021).

Efforts have been made in the domain of Arabic Natural Language Processing (ANLP) to build resources for the Arabic language. A considerable amount of research by researchers interested in ANLP has focused on building and annotating corpora to support a diverse range of ANLP tasks. The process of annotation usually involves developing guidelines for annotating Arabic corpora. These annotations serve to facilitate various NLP tasks specific to Arabic. However, it's worth noting that a majority of these studies have predominantly concentrated on creating guidelines and annotations for Modern Standard Arabic (MSA) corpora. Nevertheless, interest in dialectal Arabic corpora is steadily increasing and evolving as opportunities for dialectal studies and machine learning advancements continue to unfold.

Some notable efforts have been made in Arabic error annotation. Abuhakema et al. (2008) compiled and annotated for errors an Arabic learner corpus consisting of nine thousand words collected from texts written by students who studied Arabic as a foreign language. Alfaifi et al. (2014) worked on the Arabic Learner Corpus (ALC) project that comprises 282,000 words from students studying Arabic in Saudi Arabia. It includes written and spoken text by 942 students from 67 nationalities at pre-university and university levels in which errors are tagged following a newly created tagset. Zaghouni et al. (2014) developed an error annotation framework to annotate linguistic errors and corrections in an Arabic corpus consisting of news comments, native and non-native student essays and machine translation output. Their annotated corpus can be used to build Arabic error detection and correction tools (Zaghouni et al., 2014).

There has also been some rising interest in collecting and annotating dialectal Arabic. Habash et al. (2008) developed guidelines for the annotation of code-switching from MSA to dialectal Arabic in written text. The guidelines were developed for word and sentence level annotations of a corpus consisting of 59 documents (19K words) from newswire, web text, broadcast news and conversations. This annotated corpus can be used to develop dialect identification systems (Habash et al., 2008). Jarrar et al. (2017) compiled a corpus of more than 56, 000 Palestinian words that were annotated with rich morphological and lexical features, such as POS, stems, suffixes, prefixes, lemmas, and glosses. Furthermore, Maamouri et al. (2006) developed a Levantine Arabic treebank of conversational telephone speech and annotated 26,000 Levantine words morphologically and syntactically. Another attempt was made by Maamouri et al. (2014) who created an Egyptian Arabic treebank from informal discussion forum text and annotated it with morphological and syntactic features.

Some efforts have been made targeting the Gulf Arabic dialect. Khalifa et al. (2016) compiled and annotated the Gumar corpus which consists of 110 million words in Gulf Arabic taken from 1200 online forum novels. They used the MADAMIRA morphological tagger to morphologically tag the tokens and adapted it to cover Gulf Arabic. They annotated the corpus at the document level and added information of author's name, novel name, and dialect used in the novel. Khalifa et al. (2018) targeted the Emirati Arabic novels compiled in the Gumar corpus and annotated 212,000 words following CODA guidelines with respect to POS tags, English glosses, lemmas, dialect, and performed spelling regularisation.

Other studies related to conversation analysis have been carried out such as Elmadany et al. (2018) who were interested in speech acts in conversations. The study involved compiling an Arabic Speech-Act and Sentiment Corpus of Tweets (ArSAS) which consisted of 21,000 Arabic tweets and developing guidelines to annotate the corpus according to the speech act

and sentiment found in the tweet. The corpus can be used to build speech-act identification systems and sentiment analysis systems.

While previous efforts have laid a strong foundation for the collection and annotation of different types of corpora, our planned corpus aims to innovate and expand upon these efforts, especially in the field of ANLP and Arabic discourse studies, by developing a framework to annotate conversational Kuwaiti Arabic with conversational functions and analyse how men and women perform conversational interaction in online communication.

5.3 WhatsApp Data Annotation Methodology

5.3.1 Context of the WhatsApp Reading Club Groups

The WhatsApp reading groups that have participated in the study are three groups of mixed gender Kuwaiti participants who meet face-to-face on a monthly basis to discuss a book that they have chosen using the WhatsApp group platform. They usually meet in cafes or libraries. In these WhatsApp groups, participants choose the books they want to read, they decide on the venue for the meetings, they engage in discussions about the selected books, they share their thoughts and insights and talk about cultural or educational events happening. It is worth noting that the time frame during which this study was conducted coincided with the spread of the coronavirus (COVID-19) pandemic. The pandemic's impact on daily life, including communication patterns, cannot be understated. The pandemic, with its social distancing measures and restrictions on in-person interactions, led to a shift in how members engaged with one another. This period saw the introduction of virtual meeting platforms, such as Zoom online meetings, as an alternative means for group discussions.

5.3.2 The Development of the Conversational Function Tags and Guidelines for Conversation Annotation

In order to systematically analyse and annotate the WhatsApp reading club chats, an inductive thematic analysis was undertaken to explore the dataset and identify recurring conversational themes and patterns. The following lines will describe the steps taken to develop the tag list and the annotation guidelines.

The Tag List

The tag list was created after a rigorous thematic analysis was undertaken. The steps that are often followed in a thematic analysis are 6 main steps which are “familiarising yourself with

the data, generating initial codes, searching for themes, reviewing the themes, defining and naming the themes and finally writing up the report or manuscript” (Braun and Clarke, 2006, p.87).

In the initial interaction with the data, multiple readings of the data were conducted to become familiar with the dataset. It included skimming the entire body of data to get a preliminary sense of the chats. It was also followed by a thorough reading of the dataset and note-taking of recurring topics, patterns, terminology, and linguistic behaviour. This is the coding step of the thematic analysis which was done by actively reading the WhatsApp chats and monitoring the conversations. During this process, notes were taken next to the chat turns, tracking the topics discussed and the shifts in conversation topics. It also included observing the various linguistic conversational functions employed. In the following step, the notes that were taken were reviewed to extract themes from the conversations. The themes identified were scrutinised and an examination of how those themes were supported by examples from the text was conducted. This was done to make sure that the themes reflect the data. Then the themes were given names and descriptions. This phase underwent several iterations involving a process of refinement. This was necessary to maintain distinctive definitions of the themes and avoid any elements overlapping.

The thematic analysis of our study led to the identification of 7 distinct themes. These themes have been selected to serve as the taglist for our study. The tags and their descriptions are presented in Table. 5.1. The set of tags identified include two which are considered generic tags: Greeting and Leave-taking. These tags are of interest to us as we aim to investigate how they are utilized in online communication within WhatsApp, particularly in relation to gender differences. The remaining five tags can be categorized as procedural tags, which include Arranging Club Meeting, Social Interaction, Feedback on Club Meeting, Book Discussion, and General reading-related topics. We also intend to explore how these procedural tags are employed amongst men and women in the context of online conversations. Sec. 5.4.2 offers a comprehensive and detailed analysis of the 7 identified tags.

Conversational Function Tags	Description
Greeting	Any form of greeting as they begin the chat. e.g., Hello السلام عليكم
Leave-taking	Any reference to closure of the chat. e.g., See you نشوفكم على خير
Arranging club meeting	Any instance related to meeting arrangements: time, date, location, book choice. Any instance of users' agreement to attend the meeting or excuse for not attending. e.g., Would you like to go to a cafe to discuss? تبون نروح كافيه نتناقش؟
Book Discussion	Any discussion related to the content of the book. Any discussion related to the author of the book. Any instances where members share views/opinions about the book. If the text has a social interaction nature but a specific reference to the book, it should be tagged as book discussion. e.g., A nice novel, light and deep خوش رواية خفيفة وعميقة
Feedback on club meeting	Any feedback related to the club meeting such as if they enjoyed it, what they enjoyed about it, feedback on members' performance in the meeting, etc. e.g., We enjoyed the discussion a lot استمتعنا وايد بالمناقشة
General reading-related topics	Instances when members talk about book exhibitions. Instances where members discuss software and applications used for reading. e.g., If anyone who has a Goodreads account, please share it اللي عنده أكاونت في قود ريدز يا ريت يحط أكاونته
Social interaction	Instances of welcoming new members. Instances of new members introducing themselves. Instances of sharing personal information or social events. e.g., I'm busy with my mother at the hospital, please pray for her مشغول مع الوالدة بالمستشفى دعواتكم لها

Table 5.1 Conversational Tags and their Description

The Annotation Guidelines

Creating annotation guidelines for the WhatsApp dataset to be used by annotators when annotating the dataset was essential to ensure consistency and accuracy in the annotation process. A set of instructions on how to tag the textual segments were provided along with the description of the taglist for annotators to follow during the annotation task.

5.3.3 Recruitment and Training of Annotators

Recruitment of Annotators

Recruiting and training annotators for the annotation of the WhatsApp reading group chats was a crucial step in ensuring the accuracy and quality of the task. The annotation task was completed by a team consisting of the researcher and three English language instructors from Kuwait University, all of whom are native speakers of Kuwaiti Arabic. Their expertise, linguistic proficiency and genuine interest in language were instrumental in successfully completing the annotation of the WhatsApp dataset.

Training of Annotators

To prepare the annotators for the annotation task, three comprehensive training sessions were conducted. In these sessions, the annotators were introduced to the task in which they were asked to tag each sentence of the WhatsApp dataset, which consists of 3 WhatsApp reading club group chats, with a tag that corresponds to the sentence's conversational function using an annotation tool to facilitate the process. They were given training on how to use CATMA (Computer Assisted Text Markup and Analysis), a web-application annotation tool that is used to annotate, analyse and visualise textual data (Gius et al., 2023). They were provided with the annotation guideline sheet and tag descriptions to refer to when annotating. They were also given some annotation training on practice material. During these sessions, the annotators were given the chance to ask any questions and were asked to provide any constructive feedback on the annotation guidelines. These sessions equipped the team with the necessary knowledge and tools to annotate the WhatsApp dataset. The annotators were given a time limit to complete the annotation of the 3 WhatsApp group chats, and after completion of each group, inter-annotator agreement scores were obtained to assess the level of agreement between the annotators. Cases of disagreement were manually analysed and discussed with the annotators which led to slight modification of the guidelines.

Challenges and Annotators' Feedback

The annotators found the guidelines helpful in providing clear instructions for tagging the conversations. However, one challenge occurred when a textual segment did not neatly align with a single tag. For instance, when users in the WhatsApp groups were discussing the author of a book, it did not fit neatly under the tag `Book Discussion`. To address this, we refined the tag `Book Discussion` to include cases where the conversation was about the author.

5.3.4 The Kuwaiti Arabic Conversational Function Dataset (KACD)

The annotated WhatsApp corpus is a conversational dataset consisting of 5,289 sentences, carefully categorised and tagged with gender and conversational function information.

Conversational Function	Total No. of Sents	Gender	No. of Sents	No. of Words
Arranging Club Meeting	1840	Men	607	2782
		Women	1233	6363
Book Discussion	703	Men	411	4302
		Women	292	2788
Feedback on Club Meeting	138	Men	50	418
		Women	88	987
General Reading-related Discussion	216	Men	81	458
		Women	135	919
Greeting	274	Men	95	228
		Women	179	471
Leave-taking	3	Men	2	30
		Women	1	6
Social Interaction	2115	Men	922	5605
		Women	1193	6191
Total Number of Words				44350

Table 5.2 Descriptive Statistics for the KACD Corpus, Including Sentence and Word Counts per Tag and Gender

5.4 Results and Analysis

In this section, results of IAA scores for the four annotators are presented. This is followed by descriptive statistics of the annotated dataset. The section is concluded by reporting statistical results of the distribution of the conversational functions amongst men and women and any significant qualitative observation.

5.4.1 The Inter-Annotator Agreement Scores

	Ann 2			Ann 3			Ann 4		
	G1	G2	G3	G1	G2	G3	G1	G2	G3
Ann 1	0.77	0.74	0.48	0.69	0.81	0.51	0.90	0.91	0.93
Ann 2				0.73	0.77	0.48	0.76	0.78	0.49
Ann 3							0.67	0.89	0.53

Table 5.3 Sentence Level Inter-annotator Agreement Scores

Cohens Kappa was used to measure the Inter-Annotator Agreement (IAA). It is a statistical measure of the degree of agreement between two annotators labelling items from a given set of labels (Cohen, 1960). The IAA scores for annotating the three WhatsApp group chats by four annotators are presented in Table 5.3. Each annotator was paired with another to compute the agreement scores of their annotations.

As can be seen in Table 5.3, the overall IAA scores obtained range between moderate to near perfect agreement according to the conventional interpretation of kappa IAA scores. After comparing the IAA scores for all groups, a pattern was observed. The highest agreement among the annotators was between annotator 1 and annotator 4 (0.93) and the lowest agreement was mostly between annotator 1 and 2 (0.48). It was also noticed that 5 IAA scores were within the range of moderate agreement, 8 IAA scores were within the substantial range and 5 IAA scores were within the near perfect range.

5.4.2 Descriptive Statistics

Conversational Functions	Gender	Num.	Min.	Max.	Mean	Median	Q1	Q3	IQR
Arranging Club Meeting	Men	14	0	183	43.4	21.5	7.5	59.5	52.0
	Women	28	0	577	44.0	12.5	7.3	27.5	20.2
Book Discussion	Men	14	0	179	29.4	8.0	1.5	23.8	22.3
	Women	28	0	81	10.4	2.0	0.3	10.3	10.0
Feedback on Club Meeting	Men	14	0	20	3.6	1.0	0.0	3.8	3.8
	Women	28	0	28	3.1	0.50	0.0	3.8	3.8
General Reading-related Discussion	Men	14	0	27	5.8	3.0	0.0	7.3	7.3
	Women	28	0	32	4.8	1.0	0.0	6.0	6.0
Greeting	Men	14	0	33	6.8	1.0	0.0	7.8	7.8
	Women	28	0	75	6.4	1.5	0.0	4.0	4.0
Leave-taking	Men	14	0	1	0.14	0.0	0.0	0.0	0.0
	Women	28	0	1	0.04	0.0	0.0	0.0	0.0
Social Interaction	Men	14	0	224	66.0	35.5	4.8	133.3	128.5
	Women	28	0	284	42.6	12.5	4.3	60.5	56.2

Table 5.4 Descriptive Statistics for the 7 Conversational Functions

5.4.3 Quantitative Analysis

We were interested in studying if there is a difference between males and females in their use of conversational functions in our annotated dataset. Specifically, we wanted to study the proportion of each participant's tags that fall into the seven distinct conversational functions (Arranging Club Meeting, Book Discussion, Feedback on Club Meeting, General reading-related topics, Greeting, Leave-taking and Social Interaction). In order to capture individual differences in conversational behavior of men and women in our dataset while accounting for potential differences in the total number of utterances made by each user we performed normalisation. This was achieved by dividing the number of each participant's tags that are a specific conversational function by the total number of that participant's utterances. This method allows for a fair comparison across participants, enabling us to identify potential gender-specific patterns in the distribution of conversational functions. We tested the distribution of men and women's usage of each normalised conversational function tag using the Shapiro-Wilk normality test (sample size less than 50) to identify whether the data is normally distributed or not and to choose the statistical test to use to determine if there is a significant difference in the distribution of conversational function tags between men and women.

Tags	Gender	Shapiro-Wilk Statistic	df	Sig.
Arranging Club Meeting	Male	0.957	14	0.680
	Female	0.955	28	0.270
Book Discussion	Male	0.768	14	0.002
	Female	0.849	28	0.001
Feedback on Club Meeting	Male	0.839	14	0.016
	Female	0.626	28	0.000
General Reading-related Discussion	Male	0.754	14	0.001
	Female	0.466	28	0.000
Greeting	Male	0.824	14	0.010
	Female	0.622	28	0.000
Leave-taking	Male	0.396	14	0.000
	Female	0.188	28	0.000
Social Interaction	Male	0.926	14	0.266
	Female	0.946	28	0.160

Table 5.5 Normality Test Results for the Normalised Proportions of Tag Usage

The Shapiro-Wilk test results as shown in Table. 5.5 indicate that for the tags Arranging Club Meeting and Social Interaction the data for both male and female participants follow a normal distribution (indicated by significance values greater than 0.05). However, for the tags Book Discussion, Feedback on Club Meeting, General Reading-related Discussion, Greeting and Leave-taking the data significantly deviate from normality for both genders (indicated by significance values less than 0.05). Therefore, we apply the t-test to the usage of Book Discussion and Feedback on Club Meeting and the Mann-Whitney U test to the remaining conversational function tags.

To compare the conversational functions employed by men and in WhatsApp reading club groups to determine if there are any significant differences in their communication patterns, we formulate the following hypotheses as detailed in Table. 5.6.

Null Hypothesis (H_0)	Alternative Hypothesis (H_1)	Statistical Test
There is no difference between men and women in the proportion of utterances tagged as:	There is a difference between men and women in the proportion of utterances tagged as:	The statistical test that is used is as follows:
- Arranging Club Meeting	- Arranging Club Meeting	T-Test
- Social Interaction	- Social Interaction	T-Test
- Book Discussion	- Book Discussion	Mann-Whitney
- Feedback on Club Meeting	- Feedback on Club Meeting	Mann-Whitney
- General Reading-related Discussion	- General Reading-related Discussion	Mann-Whitney
- Greeting	- Greeting	Mann-Whitney
- Leave-taking	- Leave-taking	Mann-Whitney

Table 5.6 Hypotheses and Statistical Tests Used to Compare Men and Women Across their Usage of Conversational Functions

In our significance testing, we interpret a p-value less than 0.05 as strong evidence of a difference between men and women. If the p-value is less than 0.1, it suggests a potential difference, although it is less conclusive than the conventional threshold. When the p-value falls between 0.05 and 0.1, we report the findings at both the 0.05 and 0.1 significance levels.

Arranging Club Meeting In studying if there is a difference between men and women in their proportion of utterances tagged as Arranging Club Meeting, we performed an independent samples t-test to determine if there is a difference between the means of the two groups. The t-test yielded a p-value of 0.644 as shown in 5.7, which is greater than the conventional threshold of 0.05. Therefore, we fail to reject the null hypothesis. This suggests that there is no statistically significant difference between men and women in the proportion of utterances tagged as Arranging Club Meeting.

Conversational Function	Gender	N	Mean	SD
Arranging Club Meeting	Male	14	0.340	0.237
	Female	28	0.374	0.214
T-test P-value	0.644			

Table 5.7 T-Test Results for Arranging Club Meeting

We also conducted Mann-Whitney U test on the proportion of utterances tagged as Arranging Club Meetings since it does not assume normality to compare its result with the other tags tested using Mann-Whitney. The Mann-Whitney U test results for the Arranging Club Meeting tag in table 5.8 reveal no statistically significant difference in mean ranks between male and female participants. The mean rank for males is 19.96, and for females, it is 22.27. The Mann-Whitney U value is 174.50, with a p-value of 0.566. As the p-value exceeds the 0.05 threshold, we fail to reject the null hypothesis which means that there are no differences between men and women in the proportion of their utterances tagged as Arranging Club Meeting.

Conversational Function	Gender	N	Mean Rank	Sum of Ranks
Arranging Club Meeting	Male	14	19.96	279.50
	Female	28	22.27	623.50
Mann-Whitney U	174.50			
P-value	0.566			

Table 5.8 Mann-Whitney U Test Results for Arranging Club Meeting

Book Discussion In studying if there is a difference between men and women in their proportion of utterances tagged as Book Discussion, we performed a Mann-Whitney U Test to determine if there is a difference in the mean of the ranks between the two groups. The Mann-Whitney U test yielded a p-value of 0.084 as shown in Table 5.9, which is greater than the conventional threshold of 0.05. Therefore, we fail to reject the null hypothesis. This suggests that there is no statistically significant difference between men and women in the proportion of utterances tagged as Book Discussion.

When considering a threshold of 0.10, the p-value of 0.084 is less than this threshold. Therefore, under this criterion, we would reject the null hypothesis. This suggests that there may be some evidence of difference between men and women in the proportion of utterances tagged as Book Discussion when using a 0.10 significance level with men showing higher proportion of usage.

Feedback on Club Meeting To investigate whether there is a difference between men and women in their proportion of utterances tagged as Feedback on Club Meeting, we conducted a Mann-Whitney U Test to compare the mean ranks of the two groups. The test produced a p-value of 0.079, as indicated in Table 5.10. Since this p-value is greater than the conventional threshold of 0.05, we do not reject the null hypothesis. This indicates that

Conversational Function	Gender	N	Mean Rank	Sum of Ranks
Book Discussion	Male	14	26.07	365.00
	Female	28	19.21	538.00
Mann-Whitney U	132.000			
P-value	0.084			

Table 5.9 Mann-Whitney U Test Results for Book Discussion

there is no statistically significant difference between men and women in the proportion of utterances tagged as Feedback on Club Meeting.

However, when considering a threshold of 0.10, the p-value of 0.079 is considered less than this threshold. Therefore, under this criterion, we would reject the null hypothesis. This suggests that there may be a some difference between men and women in the proportion of utterances tagged as Feedback on Club Meeting when using a 0.10 significance level.

Conversational Function	Gender	N	Mean Rank	Sum of Ranks
Feedback in Club Meeting	Male	14	26.04	364.50
	Female	28	19.23	538.50
Mann-Whitney U	132.500			
P-value	0.079			

Table 5.10 Mann-Whitney U Test Results for Feedback in Club Meeting

General Reading-related Discussion To examine whether there is a difference between men and women in the proportion of utterances tagged as General Reading-related Discussion, we applied a Mann-Whitney U Test to compare the ranks of the two groups. The test yielded a p-value of 0.828, as shown in Table 5.11. Since this p-value exceeds the conventional threshold of 0.05, we do not reject the null hypothesis. This indicates that there is no statistically significant difference between men and women in the proportion of utterances tagged as General Reading-related Discussion.

Greeting To investigate whether there is a difference between men and women in the proportion of utterances tagged as Greeting, we conducted a Mann-Whitney U Test to compare the ranks of the two groups. The test produced a p-value of 0.860, as indicated in Table 5.12. Since this p-value is greater than the conventional threshold of 0.05, we do not

Conversational Function	Gender	N	Mean Rank	Sum of Ranks
General Reading-related Discussion	Male	14	22.07	309.0
	Female	28	21.21	594.0
Mann-Whitney U	188.000			
P-value	0.828			

Table 5.11 Mann-Whitney U Test Results for General Reading-related Discussion

reject the null hypothesis. This indicates that there is no statistically significant difference between men and women in the proportion of utterances tagged as *Greeting*.

Conversational Function	Gender	N	Mean Rank	Sum of Ranks
Greeting	Male	14	21.04	294.5
	Female	28	21.73	608.5
Mann-Whitney U	189.500			
P-value	0.860			

Table 5.12 Mann-Whitney U Test Results for Greeting

Leave-taking Leave-taking was only used 3 times in the WhatsApp group chats. Two times by men and one time by a woman. All 3 cases of leave-taking were cases of members leaving the group permanently. The lack of leave-taking expressions used may be due to the observation that “Unlike greetings, leave-takings are not perceived as a social obligation in online spheres, in the sense that a user is not reprimanded if he or she exits the chat room without mentioning something about parting from a group of users.” (Algharabali, 2010, p.104).

Social Interaction To explore whether there is a difference between men and women in the proportion of utterances tagged as *Social Interaction*, we conducted an independent samples t-test to compare the means of the two groups. The test yielded a p-value of 0.765, as shown in Table 5.13. Since this p-value is greater than the conventional threshold of 0.05, we do not reject the null hypothesis. This suggests that there is no statistically significant difference between men and women in the proportion of utterances tagged as *Social Interaction*.

Conversational Function	Gender	N	Mean	SD
Social Interaction	Male	14	0.398	0.251
	Female	28	0.377	0.190
T-test P-value		0.765		

Table 5.13 T-Test Results for Social Interaction

We also conducted a Mann-Whitney U test to compare the result with the other tags tested using Mann-Whitney. The Mann-Whitney U test results for the Social Interaction tag in table 5.14 reveal no statistically significant difference in mean ranks between male and female participants. The mean rank for males is 20.54, and for females, it is 21.98. The Mann-Whitney U value is 182.5, with a p-value of 0.719. As the p-value exceeds the 0.05 threshold, we fail to reject the null hypothesis which means there is no difference between men and women in their proportion of utterances tagged as Social Interaction.

Conversational Function	Gender	N	Mean Rank	Sum of Ranks
Social Interaction	Male	14	20.54	287.50
	Female	28	21.98	615.50
Mann-Whitney U		182.50		
P-value		0.719		

Table 5.14 Mann-Whitney U Test Results for Social Interaction

Overall, the analysis shows that we lack sufficient evidence to suggest that there is a difference between men and women in their proportion of utterances belonging to the conversational functions we studied. However, in their usage of Feedback on Club Meeting and Book Discussion we found some evidence of difference with men having a higher proportion of conversational function tags in these categories than women when considering a significance p-value of 0.10. One possible explanation for men having a higher proportion of book discussions could be their proactive or anticipatory approach to engagement. Although the discussions are meant to occur face-to-face or online during the meetings, it was noticed that men tend to express their thoughts and engage in discussions beforehand, not waiting until the scheduled meeting time. This proactive approach to book discussions may have contributed to the observed difference in proportions.

Additionally, we were interested in qualitatively analysing the language used under the conversational function tag Feedback on Club Meeting, to explore what type of feedback

is given by men and women. It was observed that both men and women maintained a positive tone when giving feedback. However, differences were noticed in the type of Arabic used. We manually examined 138 utterances tagged as `Feedback` on `Club Meeting`, consisting of 50 utterances by men and 88 by women. Our analysis revealed that 38 out of the 50 utterances by men were written in MSA, whereas only 10 out of the 88 utterances by women were in MSA.

Women also used emojis when giving feedback significantly more than men. These emojis include heart emojis, flower emojis, thumbs up emojis, clapping hand emojis and smileys. On the contrary, men used very limited emojis in type and quantity including flowers and thumbs up. Of the 138 feedback utterances examined, men used 21 emojis across their 50 utterances, accounting for an average of 0.42 emojis per utterance. Women, on the other hand, used a total of 110 emojis across their 88 utterances, averaging 1.25 emojis per utterance.

5.5 Summary

In summary, this chapter presents the compilation of the Kuwaiti Arabic Conversational Dataset (KACD), which is, to the best of our knowledge, the first publicly available dataset of Kuwaiti Arabic annotated with conversational function tags. The dataset aims to provide a valuable resource for researchers working with Kuwaiti Arabic in various linguistic and computational research.

The chapter starts by reviewing the related work on annotation studies in the field of ANLP. It then presents the process of developing the annotation framework and guidelines. It discusses the use of a thematic analysis approach to systematically analyse the data and derive meaningful conversational function tags specific to book-related discussions in Kuwaiti Arabic online conversations. Additionally, the chapter explains the process of recruiting and training annotators to perform the annotation task and the interannotator agreement scores that were calculated to create a consensus dataset.

Furthermore, we conducted a study to examine the differences in the proportions of utterances that belong to the seven conversational functions we used as tags between men and women in our dataset. This analysis provided insights into the distribution of different conversational functions within the dataset amongst men and women. Overall, the KACD represents a significant advancement in the availability of resources for Kuwaiti Arabic and sets a foundation for future research in the field.

Chapter 6

Gender Classification Using Kuwaiti Arabic Tweets:

6.1 Introduction

In chapter 4 and chapter 5, we aimed at inferring linguistic features that differentiate the language of Kuwaiti men from the language of Kuwaiti women. In order to do that, we compiled in chapter 4 a written KA dataset from WhatsApp (KAGen) and labelled it with gender information to analyse and gain insights about features that characterise the language of men and women in KA (Sec. 4.3). We tested the features inferred on a baseline gender classification system to see how well the features serve as a basis to predict gender (Sec. 4.5). We also annotated the WhatsApp dataset according to the conversational functions each sentence belongs to in chapter 5 with the aim of contributing with an annotated dataset (KACD) of conversational Kuwaiti Arabic labelled by gender and conversational function (Sec. 5.3.4). The overarching aim of this project is to test the extent to which the features inferred from the previous studies can predict gender when used in a large-scale gender classification system. This is done in this chapter using a new dataset, The Kuwaiti Arabic Twitter Dataset (KATD) which to the best of our knowledge is the first largest publicly available Kuwaiti Arabic Dataset collected from Twitter that is labelled by gender. The study conducted in this chapter uses two supervised machine learning approaches: a feature engineering approach and a deep learning approach to explore the extent to which the gender of KA Twitter users can be predicted from their tweets.

6.2 Related Work

Automatically predicting gender from text documents has been a topic of interest in many domains and for different purposes. Interest in predicting the gender of a user from their language use can help in informing marketing strategies, forensics, security measures, and advertising campaigns (Fosch-Villaronga et al., 2021; HaCohen-Kerner, 2022; Ouni et al., 2023). The task of automatically identifying the gender of a text's author has been approached in the research field as a text classification problem and an authorship detection problem (Koppel et al., 2002; Suero Montero et al., 2014).

The topic of author profiling has become a focal point in the field of Natural Language Processing and has attracted considerable attention from researchers. Early efforts in author profiling predominantly focused on the English language. Some studies targeted predicting gender from textual datasets using feature engineering approaches and different linguistic and statistical features. Rao et al. (2010) used sociolinguistic features, n-gram features, and a stacked model that combines the former two feature types in training an SVM classifier to predict the gender of users from their tweets. They noticed that the sociolinguistic features outperformed the n-gram features when tested separately. However, combining both features improved the performance of the classifier, which reached an accuracy score of 72.33%. Burger et al. (2011) built a gender classification system trained on a multilingual gender labelled dataset using text-based features from the users' profiles such as screen names, full names, bio description. Their classifiers were tested on their collected dataset and the classifier that used tweet texts only performed at 76% accuracy, while adding screen names, full names, bio descriptions, and tweets texts achieved an accuracy of 92%.

Suero Montero et al. (2014) investigated the use of emotion-based features in improving gender classification of personal journal text and blog post texts. They tested the use of BoW features, emotion-based features derived from an ontology of emotion classes and attributes, and a combination of BoW and emotion-based features with Support Vector Machines and Decision Trees. They concluded that using a combination of BoW features alongside emotion-based features achieved the highest gender classification score which reached 80% for classifying personal journal texts and over 75% for web blogs. Aravantinou et al. (2015) were interested in predicting the gender of web blog authors. They tested statistical features, POS features and language model features using 8 different machine learning classifiers. They found that language model features were the most effective features in the gender classification task. They applied a feature selection algorithm that ranks the features according to their importance and concluded that using the top 40 features with a Random Forest classifier achieves the highest accuracy result of 70.50 %.

While there are a limited number of author profiling studies for the Arabic language, this area is attracting increasing attention from researchers in ANLP. A notable contribution to research in Arabic author profiling was by Rangel et al. (2019) who presented the Author Profiling and Deception Detection Shared Task (APDA) at (PAN@FIRE2019). The dataset provided for the task (ARAP-Tweet dataset) consisted of tweets from 198 authors from 15 different Arabic dialects. More than 2000 tweets were extracted from each author and were labelled with gender and age information. There were 28 submissions addressing the Author Profiling task that aimed at predicting age, gender and language variety from tweets. The system that achieved the highest accuracy of 81% on the gender task was by MagdalenaYVino who used a combination of words and emoticons 2-grams and 3-grams as features with a Random Forest classifier (Rangel et al., 2019).

Abdul-Mageed et al. (2019) contributed to the task of gender classification in Arabic using pre-trained BERT models. They used the Arap-Tweet dataset (Zaghouani and Charfi, 2018). Their best-performing model was a multilingual BERT-based model that achieved an accuracy score of 65.31%.

In Al-Ghadir and Azmi (2019), the researchers compiled a dataset of posts labelled by gender and exported from a popular Saudi Arabian social media forum spanning the period from 2011 to 2014 to use in a gender classification system. The features they extracted were normalised lists of the k highest-scoring words and stems, ranked according to the tf-idf scoring method. They compared the use of an SVM classifier and a 1-NN (1-Nearest Neighbors) classifier. They report the highest balanced accuracy of 93.16 % which was achieved using the 1-NN classifier and setting k (highest-scoring words and stems) to 100.

As can be seen, the topic of gender classification from text has witnessed considerable progress, particularly in languages with extensive resources such as the English language. However, there is still much to be explored, especially for under-represented languages and dialects such as Kuwaiti Arabic. The work in this chapter contributes to filling this gap and provides a foundation for future research in this area.

6.3 Data Collection

- **Applying for Ethical Approval:**

In order to start the data collection of the KA tweets from Twitter, we applied and obtained ethical approval from the University of Sheffield's Ethics committee.

- **Process of Data Collection:**

The process of data collection included searching for Kuwaiti female and male participants who use Twitter and tweet in KA. This was done by creating an ad that included

a description of the study, its aims and objectives, and criteria for participation, which was posted on different social media platforms. Interested participants were asked to read and sign the information sheet and consent form included in the ad. This was then followed by manually revising all forms submitted and approving applications that were compliant with the predefined criteria which included being a Kuwaiti female or male, having a Twitter account that has their real name and photo, and tweeting in Kuwaiti Arabic. We approved 98 applications of 49 Kuwaiti female users and 49 Kuwaiti male users, all of whom willingly provided explicit consent for the collection and publication of their tweets. The tweets were retrieved in March 2023, followed by a full-archive search for all users, and extracted using Twitter’s academic API. We extracted the last 1000 tweet for each user. The compiled dataset is a balanced dataset of 49000 tweets by men and 49000 tweets by women.

- **The Compiled Dataset:**

We compiled the first largest publicly available Kuwaiti Arabic Gender-labelled Dataset (KATD). KATD comprises tweets written in KA and labelled by gender that have been extracted from Twitter which is a social media platform that is widely used among Kuwaitis, making it a prominent source of online discourse within the Kuwaiti digital community. This compiled dataset can benefit researchers and developers in the fields of Natural Language Processing and Computational Sociolinguistics as it can be used to train and evaluate machine learning models for different language related-tasks and to fill in the gap of low dialectal resources. Furthermore, researchers can use KATD to study different linguistic phenomena specific to the KA dialect in online communication, which can help in deepening insights into language variation, dynamics, and digital discourse trends within the Kuwaiti digital community. Table 6.1 presents the statistics of KATD and Fig. 6.1 illustrates the distribution of tweet lengths amongst Female and Male Kuwaiti Twitter users in our dataset.

	Female Tweets	Male Tweets
Number of Unique Users	49	49
Number of Tweets	49,000	49,000
Number of URLs	11,428	12,031
Number of Mentions	30,512	33,956
Number of Words	481,008	478,030
Total Number of Words	959,038	

Table 6.1 Descriptive Statistics of the Kuwaiti Arabic Gender-labelled Dataset (KATD)

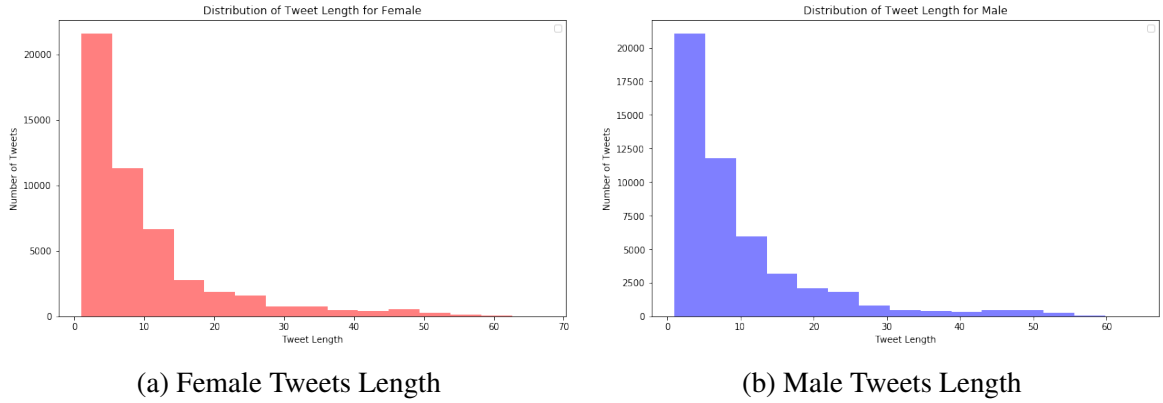


Figure 6.1 Distribution of Tweet Lengths for Females and Males

6.4 In Depth Analysis of Gender-indicative Vocabulary and Emojis

Our dataset consists of tweets written in Kuwaiti Arabic (KA) by Kuwaiti female and male users. We assumed there might be differences in the vocabulary and emojis used by men and women. Therefore, we wanted to analyse this dataset to explore these potential differences. In order to do this we followed a statistical approach. The chi-square statistical test has been used to identify characteristic vocabulary in previous studies such as in Rayson et al. (1997), in which it was used to identify words characteristic of female and male language. It was also used in Oakes et al. (2001) to identify words characteristic of document classes for document classification. We wanted to make use of this approach to extract gender characteristic vocabulary and emojis from our Twitter dataset to analyse them. We achieved this by following Oakes et al.'s implementation of the chi-square test to extract two lists from our Twitter dataset, one containing words characteristic of female language and another containing words characteristic of male language. The procedure is as follows: All tweets have been pre-processed first before creating the lists of words characteristic of male and female language. This is done because Arabic words could have multiple forms depending on where the word occurs in the sentence and what types of clitics are appended to the word.

The Chi-square test is as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad (6.1)$$

where:

χ^2 is the chi-square test statistic,

O_i is the observed frequency for each category (i) where

(i) ranges over the set of male and female categories,

E_i is the expected frequency for each category.

For each word in our Twitter dataset, we compute the observed frequencies, a , b , c , and d (see the contingency table 6.2), where a is the count of a specific word in the female tweets, b is the count of the word in the male tweets, c is the count of all other words in the female tweets, and d is the count of all other words in the male tweets.

	Female Tweets	Male Tweets
Word	a	b
\neg Word	c	d

Table 6.2 Contingency table of the observed frequencies.

The expected frequency is then computed by the following equation:

$$E_{i,j} = \frac{\text{column}_i \text{ total} * \text{row}_j \text{ total}}{\text{grand total}} \quad (6.2)$$

So, for the observed frequency a , the expected frequency $E_{1,1}$ is computed by:

$$E_{1,1} = \frac{(a+c) * (a+b)}{a+b+c+d} \quad (6.3)$$

After computing the χ^2 for each word (the sum of $\frac{(O-E)^2}{E}$ for each position in the contingency table), we then need to determine where each word belongs, i.e. is the word characteristic of male or female language? We followed Oakes et al. (2001) approach where we first obtain the ratio a/b and if this value is greater than the ratio $(a+c)/(a+d)$ we can say then that the word is more associated with the female tweet corpus, otherwise, it is considered associated with the male tweet corpus. This will result in two lists of words, one containing female characteristic words and another containing male characteristic words. Please refer to Table 6.3 for the list of female characteristic words and Table 6.4 for the list of male characteristic words extracted from KATD.

We also followed the same approach to extract emojis characteristic of men and women, and created separate lists for each gender. In the following lines, we present patterns of usage and outline themes observed and possible cultural implications:

Characteristic Vocabulary

After extracting the lists of female characteristic words and male characteristic words (Table 6.3), we scrutinised the lists and noticed that the vocabulary associated with each gender fell into certain themes, which conform to some common stereotypical patterns.

Female Characteristic Vocabulary

In Table 6.3, which presents the words characteristic of Kuwaiti female language on Twitter, we notice a variety of terms that revolve around emotions, relationships, beauty, and social expressions. The key themes we observe include:

- **Emotional and Relational Terms:**

Words like حبيبي (sweetheart/my darling), احب (I love), and *love* indicate a strong emphasis on expressing affection and emotions. Other emotional terms such as ممتنة (grateful), سعادة (happiness), and مشاعر (feelings) further emphasise the importance of expressing inner sentiments.

- **Beauty and Appearance:**

Terms related to beauty and physical appearance are prominent. Examples include جمال (beauty), حلوه (nice/pretty), تهيّل (gorgeous), and ميكب (makeup). This suggests that discussions around personal appearance and beauty standards are common among women in the dataset.

- **Religious and Cultural Expressions:**

Several words are rooted in religious or cultural expressions, such as رب (God), ماشالله (God has willed it), and الحمد لله (Thank God). This indicates the integration of faith and cultural expressions in everyday language of Kuwaiti women on Twitter.

- **Social Interactions:**

Words like *happy* and *birthday* are likely related to social celebrations and well-wishing. This indicates how Kuwaiti women engage in social interactions by celebrating birthdays and expressing these celebrations on Twitter. Terms like احبج (I love you) and عسل (honey - term of endearment) further emphasise the focus on intimate and affectionate communication.

Male Characteristic Vocabulary

Table. 6.4 which presents words characteristic of Kuwaiti male language on Twitter, reveals a distinct focus on sports, social and economic issues, politics, and group identity. Key observations include:

- **Sports and Competitions:**

Terms such as لاعب (player), دوري (league), فريق (team), and مباراة (match) indicate a significant interest in sports, particularly football. Moreover, specific references to teams and players, like مدريد (Madrid), ليفربول (Liverpool), رونالدو (Ronaldo), and برشلونة (Barcelona), suggest discussions about international sports events and figures.

- **Social and Economic Terms:**

Words related to social status and economy, such as مواطن (citizen), دينار (Dinar), and قروض (loans), are indicative of concerns related to social topics and financial matters. Furthermore, the word شركة (company) shows that topics related to business and corporate entities are also common among Kuwaiti men on Twitter. This focus suggests that males tend to discuss topics related to socio-economic status and community roles more frequently on Twitter.

- **Political Terms:**

Words like شعب (people/nation), مجلس (assembly) indicate a concern with political matters. This suggests that political discussions are more prevalent in male conversations.

- **Group Camaraderie:**

Terms like شباب (guys), ياخي (brother), and يا احبيب (dude), ياغالي (dear one) reflect a strong sense of camaraderie among Kuwaiti men on Twitter. These words indicate informal, friendly interactions often found in male social circles.

- **Emotional Expression:**

An interesting observation lies in the contrast between the language of men and women when using emotional expressions. While words associated with men, such as حزين (sad), predominantly express negative emotions, the vocabulary of Kuwaiti women in our extracted list reveals a diverse array of positive sentiments.

Word	Translation	χ^2
حبيتي	sweetheart/ my darling	377.93
احب	I love	169.41
love	-	141.94
قلب	heart	108.46
birthday	-	96.21
حب	love	92.33
happy	-	87.35
Beautiful	-	80.64
رب	God	77.57
ماشالله	God has willed it	65.43
جمال	beauty	61.19
ممتنة	grateful	45.32
سعادة	happiness	44.55
الحمدله	Thank God	43.98
حلوه	Nice/ pretty	38.45
ويع	Eww/ disgusting	37.59
تهبل	gorgeous	31.53
مشاعر	feelings	28.03
ورد	flowers	25.20
امبيه	oh my God	23.33
احبج	I love you	21.70
جميله	beautiful	19.65
حاسه	I feel	19.16
خيال	mesmerising	17.73
صالون	salon	16.95
تينن	gorgeous	13.84
امبلا	yes	9.42
قطيعه	yuck!	9.20
عسل	honey -term of endearment	7.84
ميكب	makeup	6.99

Table 6.3 Words Characteristic of Kuwaiti Female Language

Word	Translation	χ^2
مواطن	citizen	97.85
شعب	people/nation	90.86
حبيبي	dear one	88.28
لاعب	player	87.73
حبيب	nice guy	80.02
دينار	Dinar	78.46
دوري	league	71.80
مجلس	assembly	70.58
يسلم	God bless	70.06
فريق	team	68.93
لاعبين	players	57.13
رجال	men	55.94
مباراه	match	53.52
ياالحبيب	dude	52.82
ياخي	brother	52.31
حزين	sad	48.70
مدريد	Madrid	45.32
قروض	loans	44.34
ليفربول	Liverpool	43.71
ملعب	stadium	42.11
بطوله	championship	41.36
تجاره	commerce	39.82
شباب	guys	39.37
كفو	bravo	37.60
رونالدو	Ronaldo	35.16
مدرب	trainer	34.12
كاس	cup	33.54
برشلونة	Barcelona	33.16
ياغالي	dear one	32.28
شرکه	company	31.13

Table 6.4 Words Characteristic of Kuwaiti Male Language

Characteristic Emojis

Fig. 6.2 illustrates the emojis most characteristic of female and male Kuwaiti Twitter users. These emoji clouds were generated using the online tool <https://www.wordclouds.com/>. We provided the tool with emojis and their frequencies for each gender group and the tool created the visual representation. The size of each emoji corresponds to its frequency of occurrence within the gender-specific emoji list, with larger emojis indicating higher usage.

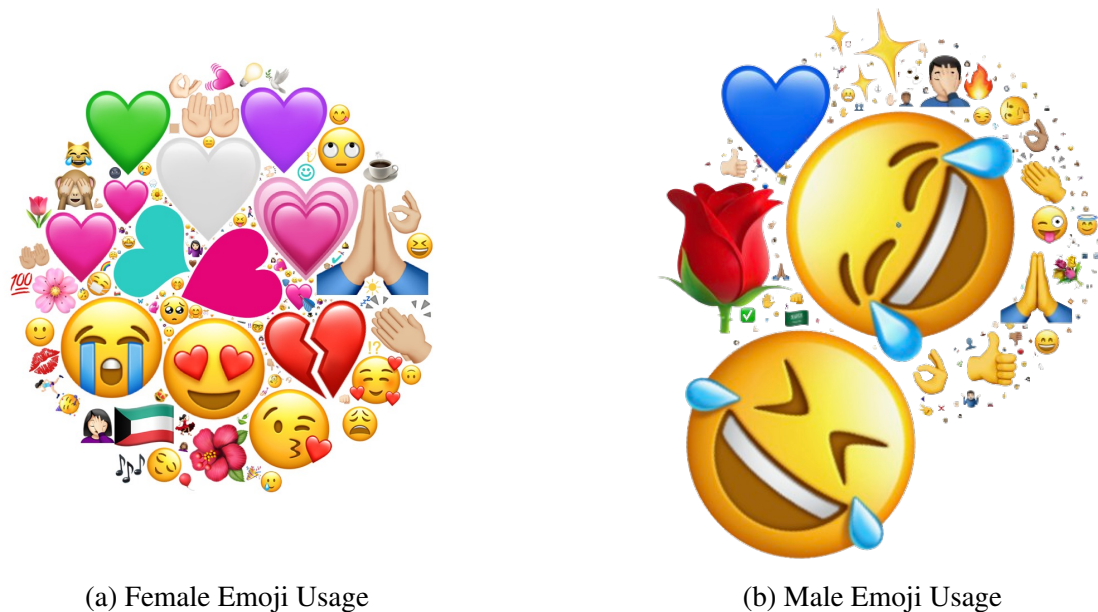


Figure 6.2 Characteristic Emojis of Females and Males Kuwaitis Twitter Users

Female-Associated Emojis

As can be seen in Fig. 6.2a, a wide range of colorful heart emojis were categorised as female emojis. These female-associated emojis include ❤️, 💜, 💖, 💚, 💕. Additionally, 💔 and 🤍 were categorised as female associated emojis which further emphasises how Kuwaiti women on Twitter practice emotional connectivity, sharing of feelings and use a wide spectrum of emojis that convey feelings from joy and love to sadness and frustration. Other commonly used emojis like 😭 and 💞 indicate an openness in expressing vulnerability and deep emotions. The emojis that were categorised as female associated emojis generally fall under themes of ‘Love and Affection’ expressed through heart emojis mentioned above and smileys like 😍, 😘, 😊 and 💋 and ‘Floral and Nature’ which includes emojis like 🌹 and 🌸, reflecting a softer, more emotive, and nurturing communication style.

Male-Associated Emojis

As for male-associated emojis in Fig. 6.2b, the emojis 😂 and 😊 were the significantly most used emojis indicated by their size in the emoji cloud. This shows how laughter-related emojis dominate the emojis Kuwaiti men use and highlights their preference for humor and light-hearted interactions. Emojis like 👍 and 🙌 were also associated with men and signify their tendency to use emojis that express approval and positive reinforcement which suggests a communicative style that values affirmation and support. Emojis related to physical activity were also associated with Kuwaiti men such as 🏃, 🚶, 🦊, 🦊, 🦊. An interesting observation was regarding the blue heart 💙 being categorised as a male characteristic emoji. This can be due to cultural and psychological factors as blue is often associated with masculinity and is neutral compared to the red heart which is strongly associated with romantic love and therefore, men may use the blue heart to express positive feelings without appearing overly sentimental. It was also observed that the 🌹 and 🌺 emojis were categorised as male characteristic emojis. This can be attributed to several factors, including traditional color perceptions and cultural stereotypes. Pink flower emojis are often viewed as more feminine due to the strong association of pink with femininity and Kuwaiti men may seek to avoid these feminine connotations, and therefore might opt for bouquet and red flower emojis, which are seen as more neutral. Overall, male characteristic emojis are limited in variety compared to female characteristic emojis and are grouped into ‘Humor and Laughter’, ‘Approval and affirmativeness’, and ‘physical activity’ which reflect a more dynamic, assertive, and energetic form of expression.

6.5 Feature Engineering Approach to Gender Classification

The first approach we adopt for our gender classification task is a feature engineering approach that integrates sociolinguistic features and statistical features. These features are derived from both the literature and our research outlined in Chapters 4 and 5. This approach is taken because it allows interpretability of our model and provides us with the opportunity to gain insights into how gender-related patterns are present in the Kuwaiti Arabic dialect. In the following subsections, we present an overview of the main steps taken in developing a supervised machine learning classifier. We then give more details about the preprocessing stage, and the feature extraction methods. We also report accuracy results of our features trained and tested on four different machine learning classifiers. Before developing our classifier, we established a baseline system using a random classifier from scikit-learn. This random classifier, applied to our dataset, achieved an accuracy of 0.50.

6.5.1 Classifier Development

This section describes the process of developing a supervised classifier for the gender classification task. There are two main phases required for the development of the classification system: training and testing. In the training phase, as shown in Fig. 6.3, the training tweets (80% of the dataset) undergo pre-processing. Then features are extracted and fed into the classifier with their corresponding gender labels (F and M) for training. The resulting trained classifier is then used for testing.

In the testing phase, shown in Fig. 6.4, the testing tweets (20% of the dataset) undergo pre-processing. Then features are extracted and fed into the trained classifier. The classifier then predicts the gender labels (F or M). We then evaluate the performance of the classifier using evaluation metrics such as accuracy.

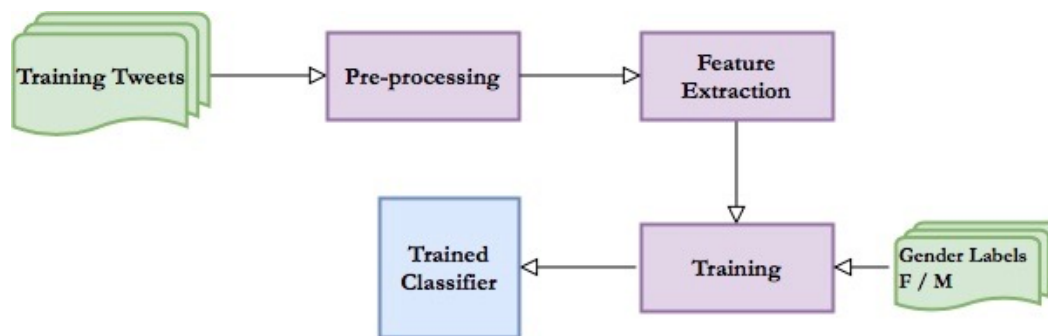


Figure 6.3 Classifier Training

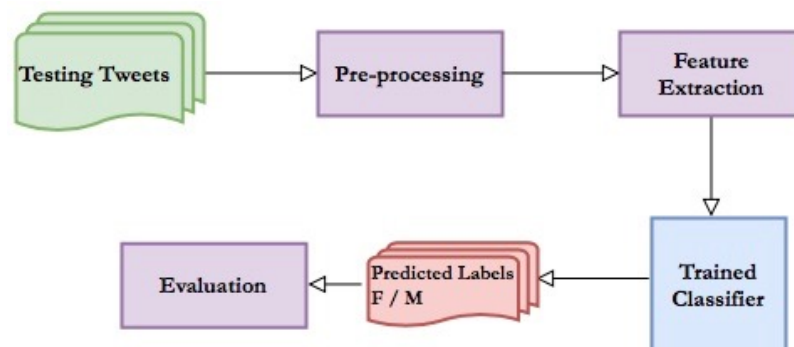


Figure 6.4 Classifier Testing

6.5.2 Tweet Pre-processing

Preprocessing the data collected before conducting feature extraction is an important step. Due to the complex and noisy nature of social media text, especially in our context of using

tweets from Twitter, that may include typos, inconsistencies in spellings of words, diacritics, URLs, and hashtags, we need to process the data to reduce noise and to convert the data to a form that is suitable for the machine learning classifiers that are used. We also wanted to test the effect of preprocessing on the performance of the gender classification system.

We first prepared the data to adhere to the ethical considerations and to maintain user privacy. We anonymised the tweets by replacing usernames and mentions with @USER. We then converted all occurrences of URLs to URL and finally tokenised the tweets using NLTK's `TweetTokenizer`, specifically designed for Twitter text. This tokeniser splits tweets into words efficiently. We then created two versions of the dataset: unprocessed (UP) and preprocessed (PP). The UP dataset underwent the previous data preparation steps only. Whereas, the PP dataset underwent the previous data preparation steps in addition to the following steps:

- **Stop Word Removal:** We removed stop words using NLTK's stop word removal method for Arabic.
- **Hashtag and Mention Filtering:** We filtered out hashtags and mentions.
- **Noise Removal:** We removed digits, links, punctuation and diacritics.
- **Normalisation:** We converted lengthened words to their normalised form.
- **Orthographic Normalisation:** We changed alef variants to `ا` and alef maksura to `ي` and teh marbuta to `ة` (common preprocessing steps for the Arabic language).

6.5.3 Feature Exploration

In this section, we present the features that we evaluate in our gender classification system. We present features including Bag of Words (BoW), stretched words, vocabulary and sentence length, punctuation marks, part of speech (POS) tags, code-switching, characteristic vocabulary, word sentiment, emojis, and URL usage, drawing from our previous studies and existing literature. Notably, BoW features and sentiment features have been explored in previous studies such as (Suero Montero et al., 2014). Additionally, stretched words, which were used as features in Chapter 4, have also been explored in the context of gender differences by (Arafat and Hamamra, 2021). Moreover, utterance length, which reflects the average length of written communication, has been studied in (Rosenfeld et al., 2016). This motivated our decision to experiment with features related to vocabulary and tweet lengths in our gender classification system. Furthermore, POS features, which capture the grammatical structure of language, were tested in combination with other features by (Aravantinou

et al., 2015), prompting us to explore their effect in classifying gender in our system as well. Emojis were studied in depth by Chen et al. (2017), who found them to be distinctive features in gender classification, which encouraged us to test how they perform in our system. Additionally, gender-specific characteristic vocabulary was analysed by (Rayson et al., 1997), revealing consistent differences in word choice between men and women. Code-switching is another feature we were particularly interested in testing, as it is commonly observed in the social media usage of Kuwaitis and may provide additional insights into gender differences. Finally, URL usage is considered a meta-related feature that could reflect differing online behaviors, and we wanted to explore its potential as a feature in our system. In the following sections, we explain how the features are implemented and we provide accuracy results for each feature tested separately using four machine learning classifiers commonly used for text classification tasks: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Decision Trees (DT). The accuracy results presented are from the validation set, which is 20% of the training data set aside for feature exploration. We experimented with different variations of the feature representation. We present the accuracy results of all feature variations to choose the variation that performs best and then use the best performing variation in our feature combination experiment.

BoW - Word Counts

We started by extracting word counts as features following a Bag-of-Word (BoW) approach. We used `CountVectorizer` from the Scikit-Learn library ¹ which converts a collection of documents into a matrix of word counts. `fit_transform()` was used to learn the vocabulary of the dataset and extract document-term matrix. The method returns the resulting matrix of word counts, where each row corresponds to a tweet and each column corresponds to a word from the unique vocabulary. The values in the matrix represent the frequency of each word in the corresponding tweet. This matrix is then fed into the machine learning classifiers. Table 6.5 shows the accuracy results of using the BoW approach on the preprocessed (PP) and unprocessed (UP) datasets. The highest accuracy score achieved was 0.64 using a Logistic Regression classifier on unprocessed tweets

¹<https://scikit-learn.org/>.

Feature Sets	SVM		KNN		LR		DT	
	PP	UP	PP	UP	PP	UP	PP	UP
Word Count	0.61	0.63	0.55	0.56	0.62	0.64	0.57	0.59

Table 6.5 Accuracy Results on Pre-processed (PP) and Unprocessed (UP) Word Count Feature

TF-IDF

Term frequency - Inverse document frequency (TF-IDF) is a commonly used feature in text classification tasks. The concept underlying TFIDF is twofold: first, a word's importance increases the more frequently it appears within a document, indicating its relevance to the document's content; second, the more documents the word appears in across some collection of documents, the less unique or discriminative the word is. TF-IDF is computed by calculating the product of a word's frequency within a particular document (tf) by its inverse document frequency across the entire document collection ($idf_{w,D}$). Consequently, words with higher TF-IDF scores are assigned higher weights, and words with lower TF-IDF scores are assigned lower weights, indicating less importance in representing the content of a given document (Sebastiani, 2002; Yun-tao et al., 2005). TF-IDF is defined as:

$$TF-IDF_{d,w,D} = tf_{d,w} \times idf_{w,D} \quad (6.4)$$

where:

- tf is the number of times a term (w) appears in a document (d).
- $idf_{w,D}$ is the number of documents in (D) that have a certain term (w).
- $idf_{w,D} = \log \frac{|D|}{df_{w,D}}$ where $|D|$ is the total number of documents.

We experimented with TFIDF values of words in our dataset as features. This was done by using `TfidfVectorizer` from Scikit-learn. This method transforms the raw text of each tweet into a numerical feature vector, where each element represents the importance of a specific word in the tweet relative to the entire corpus of tweets. The feature vector's dimensionality is equal to the size of the vocabulary of the entire tweet corpus. Each item of the vector contains the TF-IDF score of a corresponding word, computed based on its frequency in the tweet (TF) and its inverse document frequency across the corpus of tweets (IDF). Since not all words appear in every tweet, the resulting feature vectors are sparse,

with many zero values. This set of feature vectors forms a matrix-like representation of the dataset, where each row corresponds to a tweet and each column corresponds to a word in the vocabulary. These TF-IDF scores are used as features to train our machine learning models.

Feature Sets	SVM		KNN		LR		DT	
	PP	UP	PP	UP	PP	UP	PP	UP
TFIDF	0.62	0.63	0.54	0.54	0.63	0.64	0.57	0.58

Table 6.6 Accuracy Results on Pre-processed (PP) and Unprocessed (UP) TFIDF Feature

We experimented with extracting TFIDF values for both preprocessed tweets and unprocessed tweets. The accuracy results obtained using the machine learning classifiers are presented in Table. 6.6. As can be seen, the highest accuracy score achieved was 0.64 using a Logistic Regression classifier on unprocessed tweets.

Stretched Words

The identification of stretched or elongated words as a potential indicator of gender was noted in our study in sec. 4.4.2. To investigate the significance of this linguistic phenomenon in determining the gender of tweet authors, we used it as a feature in our gender classification system. The objective was to assess whether or not the presence of stretched words serves as a meaningful feature contributing to gender classification in KA. We wanted to test whether the presence of elongated words carries substantial information for gender prediction.

We considered words that had 3 and more consecutive characters as stretched words and feature extraction was conducted using two representations, a binary representation and a count representation:

1. In the binary representation, a binary value is assigned for each tweet. Specifically, a value of 0 denoted the absence of stretched words in the tweet, while a value of 1 indicated the presence of at least one stretched word within the tweet.
2. In the count representation, a value denoting the total number of stretched words found in a tweet is assigned to each tweet.

Results of using the two representations on our data showed that the two representations were very similar in their performance. The highest accuracy score achieved was 0.51 in both variations. We use the count variation when experimenting with other feature combinations.

Feature Set	Classifier			
	SVM	KNN	LR	DT
StretchedWordsBinary	0.51	0.50	0.51	0.51
StretchedWordsCount	0.51	0.50	0.51	0.51

Table 6.7 Accuracy results of classifiers using the StretchedWordsBinary and StretchedWordsCount feature sets on the unprocessed dataset.

Vocabulary and Sentence Length

We wanted to experiment with word length in tweets to see if there is a pattern associated with men and women in their use of long or short words. We tested the use of maximum word length in a tweet, and minimum word length in a tweet as features. We also wanted to test combining maximum word length with average word length, and combining minimum word length with average word length. Details of how the features are extracted are presented below:

1. Maximum Word Length:

Maximum word length is used as a feature. For each tweet, the length of each word is computed. This was done by counting the total number of characters for each word. Then the length of the longest word (maximum word length) in the tweet is extracted and used as input to the feature vector.

2. Maximum Word Length and Average Word Length:

For this feature, two values are extracted for each tweet. The first value extracted is the maximum word length. The second value extracted is the average word length. This is done by computing the sum of all word lengths in the tweet and dividing it by the total number of words in the tweet.

3. Minimum Word Length:

We also test the minimum word length as a feature. This done by counting the total number of characters for each word. Then the length of the shortest word (minimum word length) in the tweet is extracted and used as in input to the feature vector.

4. Minimum Word Length and Average Word Length:

We extract both the minimum word length and average word length for each tweet. The two values extracted from each tweet are then used as features for our gender classification system.

5. Minimum Word Length, Maximum Word Length and Average Word Length:

In this feature extraction method, we combine all features above. From each tweet, we extract the the minimum word length, the maximum word length and the average word length.

Accuracy results of the previous feature extraction methods that relied on word length are presented in Table. 6.8. As can be seen, the feature sets have been tested on pre-processed and unprocessed text using SVM, KNN, LR, and DT classifiers. The results achieved range between 0.49 and 0.53. It can also be seen that results using unprocessed text are slightly higher than using preprocessed text. We will use the (Min Word Len + Max Word Len + Avg Word Len) when experimenting with different feature combinations as this may capture more information about word length distribution amongst male and female tweets.

Feature Sets	SVM		KNN		LR		DT	
	PP	UP	PP	UP	PP	UP	PP	UP
Max Word Len	0.50	0.52	0.50	0.50	0.50	0.52	0.51	0.53
Max Word Len + Avg Word len	0.50	0.52	0.50	0.51	0.50	0.52	0.52	0.52
Min Word Len	0.50	0.52	0.51	0.49	0.50	0.52	0.51	0.52
Min Word Len + Avg Word len	0.51	0.52	0.51	0.51	0.51	0.52	0.52	0.53
Min Word Len + Max Word Len + Avg Word Len	0.50	0.52	0.51	0.51	0.50	0.52	0.52	0.53

Table 6.8 Accuracy Results on Pre-processed (PP) and Unprocessed (UP)Feature Sets

Punctuation Marks

We explored the significance of punctuation usage as a potential feature for gender classification. We experimented with three distinct variations of punctuation features: punctuation counts, counts of repeated exclamation marks, and the presence or absence of exclamation marks.

We first wanted to experiment with all punctuation symbols. In order to cover the punctuation used by Kuwaitis on Twitter, we combined both English punctuation and Arabic punctuation and saved them in a set as users may use both. We then extract for each tweet the total number of punctuation symbols used. The final feature vector contains the total counts of punctuation symbols for each tweet. This feature vector is then fed into the machine learning classifiers. The highest accuracy score achieved using this feature is 0.56 using Decision Trees as can be seen in Table. 6.9.

Feature Set	Classifier			
	SVM	KNN	LR	DT
Punctuation Count	0.55	0.55	0.55	0.56

Table 6.9 Results of Classifiers on the Punctuation Count Feature Using the Unprocessed Dataset.

In our second attempt at using punctuation as features for our gender classification system, we focused on exclamation marks as they have been observed in the literature to be indicative of gender (Waseleski, 2006). We experimented with three different representations:

1. Two or more Adjacent Exclamation Marks

In this feature representation, we identify and count occurrences of two or more consecutive exclamation marks within each tweet. We use a regular expression pattern to match sequences of two or more exclamation marks. This is done by iterating over each tweet in the input list and counting all instances of the specified pattern. The total count of consecutive exclamation marks for each tweet is appended to a feature vector and then the final feature vector for all tweets is fed into the machine learning classifiers.

2. Binary Representation of Exclamation Marks

In this feature representation, we focus on the presence or absence of exclamation marks within each tweet. For each tweet, if an exclamation mark is found, a binary value of 1 is appended to the feature vector, otherwise 0 is appended. Then the final feature vector containing binary values representing the presence or absence of exclamation marks for each tweet is fed into the machine learning classifiers.

3. Counts of Exclamation Marks

In this feature representation, instead of specifying a particular exclamation mark pattern, we tried counting the total number of exclamation marks in a tweet. For each tweet in the dataset, we count all occurrences of the exclamation marks '!'. The total count of these occurrences is then appended to the feature vector. Then the final feature vector that contains the count of exclamation marks for each tweet is fed into the machine learning classifiers.

As can be seen in Table. 6.10 the different representations of the exclamation mark had similar performance (0.50). We experiment with counts of punctuation marks in combination with different features in our feature combination experiments.

Feature Set	Classifier			
	SVM	KNN	LR	DT
Two or More Adjacent Exclamation Marks	0.50	0.50	0.50	0.50
Binary Exclamation Marks	0.50	0.50	0.50	0.50
Counts of Exclamation Marks	0.50	0.50	0.50	0.50

Table 6.10 Accuracy Results of Classifiers on Different Variations of Exclamation Mark Features

POS count

We wanted to use parts-of-speech counts (POS) as a feature for our gender classification system. We used CAMEL Tools POS tagger (the CAMElBERT implementation pre-trained on Gulf Arabic). Following the bag of words approach, this feature representation first extracts all possible Arabic POS from the text and saves them in a set. Then for every tweet, we count the occurrence of all the unique POS in that tweet. The final feature matrix representation contains columns the size of the POS tags set, each row corresponds to a tweet and includes the counts for every unique pos in the set. The POS counts are computed both for prepossessed and unprocessed text. The highest accuracy score was achieved using SVM, LR, and DT (0.54) for unprocessed text as shown in Table. 6.11.

Feature Sets	SVM		KNN		LR		DT	
	PP	UP	PP	UP	PP	UP	PP	UP
POS Counts	0.53	0.54	0.52	0.53	0.52	0.54	0.53	0.54

Table 6.11 Accuracy Results on Pre-processed (PP) and Unprocessed (UP) POS Counts

Adjective Counts in a Tweet

We experimented with counts of adjectives as features. For each tweet in the dataset, we used the CAMElBERT POS tagger to identify the adjectives present in the tweet. We count the total number of adjectives found in a tweet and append the result to our feature vector, where each item in the feature vector. The final feature vector contains the total count of adjective tags for each tweet. Table. 6.12 shows the accuracy results, where the highest accuracy result was achieved using the PP dataset with SVM, LR, and DT.

Feature Sets	SVM		KNN		LR		DT	
	PP	UP	PP	UP	PP	UP	PP	UP
Total Adjective Counts	0.52	0.51	0.50	0.51	0.52	0.51	0.52	0.51

Table 6.12 Accuracy Results on Pre-processed (PP) and Unprocessed (UP) Total Adjective Counts Feature

Code-switching

Because the linguistic phenomenon of code-switching from (KA to English) is noticed in the speech of Kuwaitis and as a written linguistic phenomenon especially in online texting, we wanted to test if this feature is associated with a certain gender. We experimented with two different feature representations:

1. Binary Representation of Code-switching

In this feature representation, our goal is to identify the presence or absence of code-switching in tweets. We process each tweet individually by iterating over it and using a regular expression to detect both Arabic and English words within each tweet. For English words, we used a regular expression that matches one or more English alphabetic characters (both uppercase and lowercase) $r'[a-zA-Z]^+$ and for Arabic words we used a regular expression that matches the unicode range for Arabic script characters $r'[\u0600-\u06FF]^+$. We implement a condition to skip over occurrences of '@USER' and 'URL' strings, as they represent mentions and URL links, which are not instances of code-switching. If the regular expression matches, indicating the presence of code-switching, we assign a binary value of 1 to the tweet; otherwise, we assign 0 to indicate the absence of code-switching. These binary values are then appended to the feature vector, where each index in the vector corresponds to a tweet in the dataset. This feature vector is then fed into the machine learning classifiers.

2. Count of Code-switching Instances

In the second feature representation, our objective is to count the instances of code-switching in tweets. We start by iterating over each tweet and using a regular expression pattern to identify sequences of English words within a tweet that has Arabic words. We consider both the presence of an English word or a sequence of English words (2 or more consecutive English words) within a tweet that contains Arabic words an instance

of code-switching. We implement a condition to exclude occurrences of ‘@USER’ and ‘URL’ strings, as mentioned earlier. For each tweet, we check if the regular expression pattern matches an English word within the tweet or an English word sequence. We count the total code-switching instances found and append the result to the feature vector. Each index in the feature vector corresponds to a tweet in the input dataset. This feature vector, containing the counts of code-switching instances for each tweet is then used as an input for the machine learning classifiers.

Feature Set	Classifier			
	SVM	KNN	LR	DS_Tree
Binary Code-switching	0.50	0.50	0.50	0.50
Count of Code-switching Instances	0.52	0.52	0.52	0.52

Table 6.13 Accuracy Results of Classifiers on Different Variations of Code-switching Features

Accuracy results of using the two different representations of the code-switching feature are presented in Table. 6.13. As can be seen the accuracy scores range in the 50s across different classifiers. The count of code-switching instances are slightly higher (0.52) than the binary representation (0.50). This suggests that the method of quantifying code-switching instances provides a slightly improved accuracy in distinguishing between tweets with and without code-switching occurrences and therefore will be used when using different feature combinations.

Word Embedding

Word embedding or distributed word representation is a technique that is commonly used in text classification tasks due to its ability to capture semantics of words. This technique represents words as dense vectors in a high-dimensional space where similar words have similar vectors (Almeida and Xexéo, 2019). We wanted to use word embeddings as features for our gender classification system. We used ArWordVec built by Fouad et al. (2020) which comprises a set of pre-trained word embeddings trained on 55 million Arabic tweets covering a wide range of topics. ArWordVec uses two main approaches to building the word embeddings: word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). Word2vec supports continuous bag-of-words (CBOW) and Skip-gram (SG).

We performed feature extraction for word embeddings using the ArWordVec (Word2Vec - CBOW) model, specifically the model CBOW-500-3-400 (vector size= 500, window size=3).

For each tweet in the input text, we tokenise the text by splitting it into individual words. We then calculate the sentence embedding for each tweet by averaging the word embedding vectors of its constituent tokens, representing the tweet as a dense vector. Finally, a NumPy array is returned containing the sentence embeddings for all sentences in the input text. Each row of the array corresponds to the embedding of a single sentence.

Feature Sets	SVM		KNN		LR		DT	
	PP	UP	PP	UP	PP	UP	PP	UP
Word Embeddings	0.60	0.60	0.58	0.57	0.60	0.60	0.56	0.55

Table 6.14 Accuracy Results on Pre-processed (PP) and Unprocessed (UP) Word Embeddings

Table 6.14 presents accuracy results of using word embeddings as features for our gender classification system. As can be seen, the highest accuracy scores were achieved using an SVM and a Logistic Regression classifier (0.60) on both preprocessed and unprocessed text.

Characteristic Vocabulary

The results of the in depth analysis of gender-indicative vocabulary and emojis in Sec. 6.4 revealed interesting observations amongst females and males and could be a good indicator of the gender of the tweets' authors. Therefore, the same approach was used to create two lists of characteristic vocabulary for females and males from the training data to be used as features. We implement two variations for this feature extraction method:

- **Variation 1:** for each tweet, we count the number of words that occur in the list of female characteristic words, and the number of words that occur in the list of male characteristic words. We then append these two values to the feature vector.
- **Variation 2:** In this variation, the count of words appearing in the male characteristic word list is given a negative value and the count of words appearing in the female characteristic word list is given a positive value. Then, the counts from both lists are summed to obtain a total score for each tweet, and this score is appended to the feature vector.

Table 6.15 shows the accuracy results using the two variations explained above on the PP and UP datasets. The results are highly similar, but variation 1 is slightly higher, so it will be used for further experiments.

Feature Set	SVM		KNN		LR		DT	
	PP	UP	PP	UP	PP	UP	PP	UP
Variation 1	0.61	0.61	0.56	0.56	0.61	0.61	0.61	0.61
Variation 2	0.60	0.60	0.51	0.51	0.60	0.60	0.60	0.60

Table 6.15 Accuracy Results of Classifiers on Different Variations of Characteristic Vocabulary Features (Pre-processed (PP) and Unprocessed (UP))

Word Sentiment

We wanted to explore the potential of sentiment lexica for gender detection of users in our Twitter dataset. Investigation of this feature was prompted by observations from the previous section related to gender characteristic vocabulary where a pattern was noticed: men predominantly employ negative emotion words, while women favour positive emotion words in their online interactions on Twitter. We wanted to explore the interplay between language, sentiment, and gender and test whether sentiment lexica can serve as an effective feature for discerning the gender of users in our Twitter dataset.

- Arabic Sentiment Lexicon:

We used two publicly available Arabic sentiment lexica from Mohammad et al. (2016): Arabic Emoticon Lexicon, and Arabic Hashtag Lexicon (dialectal).

1. Arabic Emoticon Lexicon:

This lexicon was generated through collecting nearly one million Arabic tweets that contained emoticons ‘:)’ which were considered positive indicators or ‘:(’ which were considered negative indicators. To create the lexicon they chose the words that appeared at least 5 times in the tweets. For these words, they extracted 3 scores: positive occurrence count (how many times the word appeared in a tweet that has a positive indicator), negative occurrence count (how many times the word appeared in a tweet that has a negative indicator) and sentiment score which is calculated by subtracting the word’s association with negative emoticons from its association with positive emoticons through Pointwise Mutual Information (PMI): $\text{SentimentScore}(w) = \text{PMI}(w, \text{pos}) - \text{PMI}(w, \text{neg})$.

2. Arabic Hashtag Lexicon (Dialectal)

This lexicon was compiled following the same procedure followed in compiling the Arabic Emoticon Lexicon explained above. However, they depended on a

sentiment seed lexicon of 483 dialectal Arabic sentiment seed words from Twitter built by Refaee and Rieser (2014) to create the Arabic Hashtag Lexicon . These sentiment seed words were used to extract tweets to build the Dialectal Arabic Hashtag Lexicon and computing three scores (average positive occurrence count, average negative occurrence count and average sentiment score) for each tweet.

We used both the Arabic Emoticon Lexicon and the Dialectal Arabic Hashtag Lexicon to extract word sentiment scores as features for our gender classification system. For each tweet in our dataset, we first tokenised the tweet, and then searched the Dialectal Arabic Hashtag Lexicon for the token, if not found we search the Arabic Emoticon Lexicon and we compute the average positive occurrence count, average negative occurrence count and average sentiment score for all tokens in the tweet and use these scores as features.

- Sentence Sentiment Using CAMeLTools:

We used CAMeLBERT-DA sentiment analysis model from CAMeLTools that was built by fine-tuning CAMeLBERT Dialectal Arabic (DA) model using ASTD, ArSAS, and SemEval datasets (Inoue et al., 2021b). Given a sentence, CAMeLBERT-DA SA model returns the sentiment label of the sentence (positive, negative or neutral). For each tweet in our dataset, we extracted the sentiment using CAMeLBERT-DA SA. The final feature vector contains sentiment labels corresponding to each tweet. This feature vector is then fed into our gender classification system.

- Sentence Sentiment Using the Mazajak Online Tool:

We also used Mazajak (an online Arabic sentiment analyser) which was built using a convolutional neural network (CNN) followed by a long short-term memory (LSTM) (Farha and Magdy, 2019). We used the online SA tool Mazajak through its API. Mazajak has been trained on the SemEval and ASTD datasets combined. We sent the list of tweets to Mazajak through its API and each tweet was processed individually, and the sentiment label for each tweet is then returned, categorising it as positive, negative, or neutral. Once the sentiment labels are obtained for all the tweets, they are used as features for our gender classification system.

Feature Sets	SVM		KNN		LR		DT	
	PP	UP	PP	UP	PP	UP	PP	UP
Sentiment Lexicon	0.50	0.51	0.52	0.52	0.50	0.51	0.52	0.53
Sentiment CamelTools	0.50	0.52	0.50	0.50	0.50	0.52	0.50	0.52
Sentiment Mazajak	0.51	0.52	0.51	0.52	0.51	0.52	0.51	0.52

Table 6.16 Accuracy Results on Pre-processed (PP) and Unprocessed (UP) Sentiment Analysis Scores.

Table 6.16. presents the accuracy scores of each sentiment analysis feature set across different classifiers. The table compares the performance of the sentiment lexicon, Sentiment CamelTools, and Sentiment Mazajak feature sets using four different classifiers for both pre-processed (PP) and unprocessed (UP) text. We will be using the sentiment lexicon feature in our feature combinations due to it performing slightly better than the other two feature sets.

Emojis

Social media users frequently incorporate emojis into their textual messages or tweets as a means of expressing emotions and opinions. Our study in Sec. 4.4.2 has uncovered associations between patterns of emoji usage and gender and building on these findings, we aim to use emojis as features within our gender classification system. To achieve this, we have explored and tested several variations of the emoji feature, aiming to identify the most informative representation. The following sections detail the various variations of the emoji feature that we have investigated.

- **Emoji Unigram**

Following the bag of words approach, this representation extracts unique emojis from the text and saves them in a set. Then for every sentence, we count the occurrence of all the unique emojis. The final feature matrix representation contains vectors, each vector corresponds to a tweet and includes the counts for every unique emoji in the set. The highest accuracy score was achieved using decision trees (0.64) as can be seen in Table. 6.17.

- **Emoji Bigram**

Similar to emoji unigram, this representation extracts all possible emoji bigrams in the dataset and saves them in a set. Then for each tweet, we count the occurrence of all the

unique emoji bigrams. The final feature matrix representation contains vectors, each vector corresponds to a tweet and includes the counts for every unique emoji bigram in the set. The highest accuracy score was achieved was 0.54 in all classifiers as can be seen in Table. 6.17.

- **Textual Emoji Representation**

This approach extracts all emojis from text and uses the `demojize()` function from Emoji library that converts emojis to their textual names. Then, it uses `CountVectorizer` function from Scikit-learn to get the counts of these textual emojis. The method returns a matrix of the textual emojis and their corresponding counts. The columns are the textual emoji names and the rows correspond to the tweets, where each row represents the counts of the textual emoji names found in the tweet. This achieved an accuracy score of (0.50) across all classifiers as can be seen in Table. 6.17.

- **Emojis Characteristic of Gender**

In our feature extraction method, we use the chi-square test following Oakes et al. (2001)'s implementation to extract two lists from our Twitter training dataset, one containing female characteristic emojis and another containing male characteristic emojis. The Chi-square test used to create these two lists is explained above in Sec. 6.4. Then, for each tweet in our dataset, we extract the emojis and append two values to the feature vector: one corresponding to the count the emojis found in the female characteristic emojis list and the second corresponding to the count of the emojis found in the male characteristic emojis list. The highest accuracy score was achieved was 0.61 in using SVM, LR, and DT as can be seen in Table. 6.17.

- **Emoji Sentiment**

Emojis are extensively used by online users to convey emotions and reactions. Embedded with sentiment, emojis may be potential indicators of gender. Saif M. Mohammad and Kiritchenko (2016) created a dictionary of emojis and their corresponding polarity scores. We wanted to test if sentiment of emojis are features indicative of gender in our gender classification system. We did this by looping over the tweets and for each emoji found in the tweet, 3 scores were given pertaining to how much positive, negative and neutral sentiments the emoji contains. We computed the average of all positive scores, negative scores and neutral scores of all emojis used in a sentence and used these scores as features fed into the machine learning model. Table 6.17 shows the accuracy of this feature using four different machine learning classifiers. The highest score achieved was using DT. It achieved an accuracy score of 0.60.

Feature Set	Classifier			
	SVM	KNN	LR	DT
Emoji Unigram	0.63	0.62	0.63	0.64
Emoji Bigram	0.54	0.54	0.54	0.54
Textual Emoji Representation	0.50	0.50	0.50	0.50
Emojis Characteristics of Gender	0.61	0.52	0.61	0.61
Emoji Sentiment	0.55	0.53	0.55	0.60

Table 6.17 Classifier Accuracy for Different Emojis Features

Table 6.17 shows the accuracy results for the different emoji features on the UP data. The emojiUnigrams feature set achieved the highest accuracy across all classifiers, with an accuracy range of 0.62 to 0.64. Therefore, this feature set will be used in combination with other features when exploring various feature combinations for gender classification.

URL Usage

We also wanted to explore the use of URL links as potential features to investigate the relationship between the gender of Kuwaiti Twitter users and the sharing of URL links on the platform. During the preprocessing stage, URL links in tweets were replaced with the string URL to facilitate further analysis.

In our feature extraction method, we iterated over each tweet and assigned a binary value to indicate the presence or absence of a URL link. Tweets without any URL links were assigned a value of 0, while tweets containing URL links were assigned a value of 1. This binary feature extraction process was applied to the unprocessed tweets, as the "URL" strings had been removed during preprocessing.

The resulting feature vector comprised binary values corresponding to each tweet, representing the presence or absence of URL links. The highest accuracy score achieved was 0.51 using SVM, LR, and DT as shown in Table. 6.18.

Feature	Classifier			
	SVM	KNN	LR	DT
URL	0.51	0.50	0.51	0.51

Table 6.18 Accuracy Results of Classifiers on URL Usage

In summary, Table 6.19 presents a summary of the performance results for the various features and their variations used in our gender classification system. For each feature/feature variation, the table lists the range of accuracy achieved, the best-performing classifier, as well as whether the highest score was achieved with pre-processed data or unprocessed data.

Word count, TF-IDF, and emoji unigrams each achieved the highest accuracy of 0.64 on their own, which shows their effectiveness in capturing gender-related patterns. The performance of the word count feature shows noticeable differences in word usage between genders in Kuwaiti Arabic, possibly due to differences in verbosity. TF-IDF's accuracy highlights the importance of specific words common in one gender's tweets but rare overall, which may reflect gender-specific topics or language. Similarly, emoji unigrams outperformed other features, showing that men and women use emojis differently to express emotions or social cues. Word embeddings and characteristic vocabulary, with accuracies of 0.60 and 0.61 respectively, demonstrate that the semantic content of words and the presence of gender-specific vocabulary contribute in distinguishing male and female communication patterns. Conversely, features like punctuation, stretched words, and code-switching provided limited improvement, with accuracies around 0.50 to 0.52, which shows that these features may be less prominent in gender distinctions observed in Kuwaiti Arabic tweets.

6.5.4 Feature Selection Methods

After conducting a feature exploration experiment, we proceeded with feature selection to identify the most informative ones for our classification task. This process is essential for focusing on the features that significantly impact model performance while reducing dimensionality. We evaluated the selected features using 10-fold cross-validation. The use of 10-fold cross-validation also allowed us to prepare for subsequent statistical analysis in sec. 6.7.1, where we compared the models to determine if there were significant differences in their performance.

Experiment 1: Using Mutual Information for Feature Selection

We wanted to experiment with using a feature selection method to test how the informative features would perform if combined together. We used Mutual information (`mutual_info_classif`) from Scikit-learn for feature selection. This method uses non-parametric techniques that depend on estimating entropy from k-nearest neighbors distances (Pedregosa et al., 2011). The idea behind it is measuring the dependency between two variables by quantifying the amount of information obtained about one variable through the other. The method assigns higher scores to features found informative in predicting the target variable and lower scores

Feature	Accuracy	Best Classifier	PP?
BoW			
Word Counts	0.55 - 0.64	LR	UP
TF-IDF	0.54 - 0.64	LR	UP
Stretched Words			
StretchedWordsBinary	0.50 - 0.51	SVM/ LR/ DT	UP
StretchedWordsCount	0.50 - 0.51	SVM/ LR/ DT	UP
Vocabulary and Sentence Length			
Max Word Len	0.50 - 0.53	DT	UP
Max Word Len + Avg Word len	0.50 - 0.52	SVM/ LR/ DT	UP
Min Word Len	0.49 - 0.52	SVM/ LR/ DT	UP
Min Word Len + Avg Word len	0.51 - 0.53	DT	UP
Min Word Len + Max Word Len + Avg Word Len	0.50 - 0.53	DT	UP
Punctuation Marks			
Punctuation Count	0.55 - 0.56	DT	UP
Two or More Adjacent Exclamation Marks	0.50	SVM/ KNN/ LR/ DT	UP
Binary Exclamation Marks	0.50	SVM/ KNN/ LR/ DT	UP
Counts of Exclamation Marks	0.50	SVM/ KNN/ LR/ DT	UP
POS			
POS Counts	0.52 - 0.54	SVM/ LR/ DT	UP
Total Adjective Counts	0.50 - 0.52	SVM/ LR/ DT	PP
Code-switching			
Binary Code-switching	0.50	SVM/ KNN/ LR/ DT	UP
Count of Code-switching Instances	0.52	SVM/ KNN/ LR/ DT	UP
Word Embedding			
Word Embedding	0.55 - 0.60	SVM/ LR	Both
Characteristic Vocabulary			
Variation 1	0.56 - 0.61	SVM/ LR/ DT	Both
Variation 2	0.51 - 0.60	SVM/ LR/ DT	Both
Word Sentiment			
Sentiment Lexicon	0.50 - 0.53	DT	UP
Sentiment CamelTools	0.50 - 0.52	SVM/ LR/ DT	UP
Sentiment Mazajak	0.51 - 0.52	SVM/ KNN/ LR/ DT	UP
Emojis			
Emoji Unigrams	0.62 - 0.64	DT	UP
Emoji Bigrams	0.54	SVM/ KNN/ LR/ DT	UP
Textual Emoji Representation	0.50	SVM/ KNN/ LR/ DT	UP
Emojis Characteristics of Gender	0.52 - 0.61	SVM/ LR/ DT	UP
Emoji Sentiment	0.53 - 0.60	DT	UP
URL Usage			
URL	0.50 - 0.51	SVM/ LR/ DT	UP

Table 6.19 Summary of Features Performance Results.

to less informative features. Another aim of using Mutual information is that it aids in reducing the dimensionality of the feature space while preserving relevant information for classification (Vergara and Estévez, 2014).

In our context, we use (`mutual_info_classif`) to calculate the mutual information between each of our features presented in Sec. 6.5.3 and the target variable (male or female). The method returns a zero value for a feature that is considered independent from the target variable and assigns higher values for informative features. In our experiment, we applied mutual information on 11 of our engineered features (POS, emoji unigrams, sentiment lexicon features, punctuation counts, total adjective counts, vocabulary and sentence length, characteristic vocabulary, stretched words, URL counts, code-switching, and exclamation marks). We excluded the TFIDF feature and word embeddings because processing these features was computationally intensive so we chose to experiment with using only features that have non-zero mutual information scores and then adding TFIDF and word embeddings to them in different feature combinations in our gender classification system to see how well they contribute to predicting the gender of the Kuwaiti Twitter users.

The returned features with non-zero values (ranked from highest to lowest mutual information scores) are 10 features: sentiment lexicon features, characteristic vocabulary, code-switching, vocabulary and sentence length, POS, emoji unigrams, punctuation counts, exclamation marks, total adjective counts, and stretched words.

We iteratively eliminated features from the set, beginning with the lowest-ranking ones based on the ordered feature list obtained from the `mutual_info_classif` function. At each step, we assessed the model's accuracy to determine the impact of the removal. This process continued until a drop in accuracy was observed, indicating that the discarded features were important for the model's performance. Our goal was to identify the smallest combination of features that achieves the highest accuracy.

Table. 6.20 shows the accuracy results of using the combination of the former feature combination on PP and UP tweets using four different machine learning classifiers. The highest accuracy score achieved was 0.68 using an SVM classifier on unprocessed tweets. The experiments showed that removing up to 3 of the features from the bottom of the list (exclamation marks, total adjective counts, and stretched words) had no effect on the accuracy score. However, removing the fourth feature from the bottom of the feature list (punctuation counts) reduced the accuracy score. So, for further experiments of feature combination, the 'Top 7 Features' are going to be used.

Feature Sets	SVM		KNN		LR		DT	
	UP	PP	UP	PP	UP	PP	UP	PP
All Non-zero Features	0.68	0.67	0.62	0.60	0.67	0.67	0.61	0.60
Top 9 Features	0.68	0.67	0.62	0.60	0.67	0.67	0.61	0.60
Top 8 Features	0.68	0.68	0.62	0.60	0.67	0.67	0.61	0.60
Top 7 Features	0.68	0.67	0.62	0.60	0.67	0.67	0.61	0.60
Top 6 Features	0.67	0.67	0.60	0.60	0.67	0.66	0.60	0.60

Table 6.20 Accuracy Results of Different Feature combinations using Mutual Information on Unprocessed (UP) and Preprocessed (PP) Tweets

Experiment 2: Using ANOVA F-test for Feature Selection

We also tried another feature extraction method from Scikit-learn: `f_classif`, which uses the ANOVA F-test and used it to determine the most informative features for classifying gender based on the tweets. The ANOVA F-test is a statistical test used to identify whether there are significant differences between the means of two or more groups. It assesses the relationship between each feature and the target variable, gender, in our case (Elssied et al., 2014). The method takes in a feature matrix and its corresponding tags and returns f-scores which indicate the ratio of variance between the groups to the variance within the groups, and their associated p-values which assess the significance of this ratio. We used the f-scores to rank the features. The higher the f-score, the more significant the feature is in distinguishing between the gender of users. We selected the features with p-values less than 0.05 to be considered statistically significant. The feature list with p-values less than 0.05 includes characteristic vocabulary, code switching, total adjective count, stretched words, and URL. Table 6.21 shows the accuracy results of these selected features. As can be seen, the accuracy results of the feature combination chosen based on the ANOVA F-test feature selection method were much lower compared to the feature combination chosen using Mutual Information. Further feature combination experiments will use the feature combination obtained using Mutual Information.

Feature Sets	SVM		KNN		LR		DT	
	UP	PP	UP	PP	UP	PP	UP	PP
Features with p-value < 0.05	0.61	0.61	0.57	0.58	0.61	0.61	0.60	0.61

Table 6.21 Comparison of Accuracy Scores for Features with P-Values Less than 0.05 According to ANOVA F-test Feature Selection Method Using Unprocessed (UP) and Preprocessed (PP) Tweets

6.5.5 Combined Feature Experiments

In this section, we experiment with different feature combinations to identify the optimal set that enhances the performance of our gender classification model. The combinations are selected based on the individual performance of features, with the hypothesis that certain combinations may lead to better classification outcomes. We evaluate these combinations using 10-fold cross-validation to ensure that the selected features contribute effectively across different data splits. The specific combinations tested are detailed below:

Combination 1: Top 7 + TFIDF As TFIDF was one of the best performing features when tested on its own, we wanted to experiment with combining it with the feature set obtained through the mutual information feature selection method. Therefore, we tested combining TFIDF with the top 7 features which are:

1. Sentiment lexicon features: Indicators of the emotional tone in the tweets.
2. Characteristic vocabulary: Specific words or phrases unique to certain genders.
3. Code-switching: The use of English and Arabic within tweets.
4. Vocabulary and sentence length: word length and sentence length to capture syntactic structure.
5. POS: Grammatical structure and usage patterns.
6. Emoji unigrams: Usage patterns of individual emojis.
7. Punctuation counts: Frequency of punctuation marks, which can indicate writing style.

By integrating TFIDF with these features, we aim to improve the model's performance by combining the strengths of term importance with different linguistic and syntactic features.

Combination 2: Top 7 + Word Embeddings Word Embeddings performed well individually compared to other features tested. Therefore, we experimented with combining WE with the top 7 features listed above. We aimed to test if combining features that capture semantic understanding with linguistic and syntactic information can improve the classification performance.

Combination 3: Top 7 + TFIDF + Word Embeddings Because TFIDF can capture the importance of words, and WE can provide semantic context, and the top 7 features can add diverse linguistic insights, we wanted to combine all these features and test whether this combination has a positive effect on the performance of the classification system.

Combination 4: TFIDF + Word Embeddings + Emoji Unigrams In this combination, we combine the best-performing individual features: TFIDF, Word Embeddings, and emoji unigrams. Each of these features achieved high accuracy scores independently. By combining them, we want to test if their collective use can improve the model's performance.

Combination 5: All Features This combination includes all the engineered features: POS, emoji unigrams, sentiment lexicon features, punctuation counts, total adjective counts, vocabulary and sentence length, characteristic vocabulary, stretched words, URL counts, code-switching, exclamation marks, TFIDF, and Word Embeddings. We combined all 13 features to examine how well the full feature set performs in of our gender classification model. This combination aims to capture as much information as possible from various linguistic, syntactic, and semantic aspects of the tweets.

Feature Sets	SVM		KNN		LR		DT	
	UP	PP	UP	PP	UP	PP	UP	PP
Comb. 1: Top 7 + TFIDF	0.69	0.67	0.62	0.60	0.68	0.68	0.62	0.61
Comb. 2: Top 7 + WE	0.68	0.67	0.61	0.61	0.68	0.67	0.59	0.58
Comb. 3: Top 7 + TFIDF + WE	0.69	0.67	0.61	0.61	0.68	0.67	0.59	0.59
Comb. 4: TF-IDF + emoji unigrams + WE	0.67	0.67	0.59	0.59	0.67	0.67	0.59	0.59
Comb. 5: All Features	0.69	0.67	0.61	0.62	0.68	0.68	0.60	0.60

Table 6.22 Comparison of Accuracy Scores for Different Feature Combinations

As seen in Table. 6.22, the top 7 features with TFIDF (Combination 1) and using all features (Combination 5) both achieve the highest accuracy with SVM (0.69) on UP text. This

suggests that TFIDF and the selected top features are the most effective features in achieving high accuracy in our gender classification task. Other combinations did not significantly improve the results. Further analysis is presented in Sec. 6.7.

6.6 Deep Learning Approach to Gender Classification

Given the surge in interest surrounding pre-trained large language models, this section adopts a deep learning approach to assess the performance of pre-trained language models in predicting the gender of Kuwaiti Arabic Twitter users. To accomplish this, we fine-tuned three transformer-based models, each of which adheres to the BERT architecture (Devlin et al., 2018a).

Fine-tuning a transformer-based model involves exposing the pre-trained model to our training data, essentially further training the pre-trained model on our specific dataset. This process begins with tokenising the tweets, followed by the addition of special token [CLS] at the beginning of the sequence (tweets in our case) as illustrated in Figure 6.5. Subsequently, this results in a sentence embedding that is fed into the model.

During training, the pre-trained model learns the weights of the model by computing the loss between the true labels and the predicted labels using cross-entropy, followed by updating the weights. To generate the predicted labels, the output vector (\mathbf{y}_{CLS}) of the [CLS] token of the final layer is passed through a classifier head consisting of a dense layer followed by a softmax function. The weights are then multiplied by (\mathbf{y}_{CLS}) and passed through the softmax function. This is shown in equation 6.5, from Jurafsky and Martin (2023, p.253).

$$y = \text{softmax}(\mathbf{W}_{CY_{CLS}}) \quad (6.5)$$

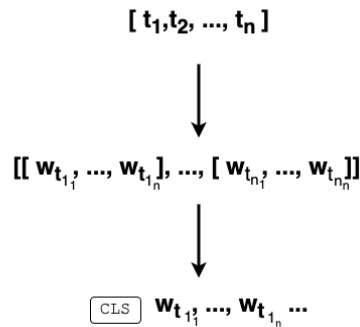


Figure 6.5 Example Input Sequence to a Transformer Based Model.

6.6.1 Models and Hyper-parameters

Three transformer based models were fine tuned for our gender classification tasks. These models are:

- **CAMeLBERT**: CAMeLBERT is a set of BERT models designed for Arabic text. It includes models for Modern Standard Arabic (MSA), dialectal Arabic (DA), and classical Arabic (CA), as well as a combined model. There are also smaller versions of the MSA model. We used `bert-base-arabic-camelbert-da` model that was pre-trained on dialectal Arabic (Inoue et al., 2021a).

Hyper-parameters: we fine-tuned this model using a learning rate of $1.44E - 05$, and it was trained for 3 epochs. The maximum sequence length was set to 128 tokens.

- **Marbert**: is a large-scale language model designed for both Dialectal Arabic (DA) and Modern Standard Arabic (MSA). MARBERT, was trained using a diverse dataset of 1 billion Arabic tweets from their extensive repository of approximately 6 billion tweets. Tweets containing at least 3 Arabic words were selected, regardless of the presence of non-Arabic text. This dataset amounts to 128GB of text, equivalent to 15.6 billion tokens (Abdul-Mageed et al., 2021).

Hyper-parameters: this model was fine-tuned using a learning rate of $7.658062e - 06$, and it was trained for 3 epochs. The maximum sequence length was set to 128 token.

- **QARiB**: or QCRI Arabic and Dialectal BERT, is a language model trained on a vast dataset of approximately 420 million tweets and 180 million sentences. The tweets were gathered using the Twitter API with a language filter (`lang:ar`), while the text data came from sources such as Arabic GigaWord, Abulkhair Arabic Corpus, and OPUS (Abdelali et al., 2021).

Hyper-parameters: this model was fine-tuned using a learning rate of $1.05101e - 07$, and it was trained for 5 epochs. The maximum sequence length was set to 128 token.

We manually experimented with various hyperparameter values, including learning rates, number of epochs, and maximum sequence lengths. The maximum sequence lengths was set 128 to cover all tweets without losing information as the longest tweet in our dataset has 114 tokens. In addition, the function `lr_find()` from `ktrain` was used to help find the best learning rate possible for each model.

The implementation of the deep learning approach experiments was carried out using the `ktrain` library. This is a python tool that wraps up other deep learning libraries, e.g. TensorFlow (Maiya, 2020). To evaluate the models, we employed 10-fold cross-validation on Pre-processed (PP) and Unprocessed (UP) text. Results are presented in Table 6.23.

6.6.2 Accuracy Results Using Transformer Models

In this section, we evaluate the performance of several Transformer models mentioned above on the task of gender classification of KA tweets. We provide a detailed comparison of accuracy results achieved by different models. The table below summarises the classification accuracy of each Transformer model used in our experiments:

Pre-Trained Transformer Models	Accuracy	
	PP	UP
CAMeLBERT	0.53	0.72
Marbert	0.52	0.73
QARiB	0.56	0.66

Table 6.23 Accuracy Results on Pre-processed (PP) and Unprocessed (UP) With Transformer Models.

6.7 Analysis and Discussion

This section provides an analysis and discussion of the results obtained from the feature engineering approach and the deep learning approach employed in this study. Each approach is examined thoroughly and a comparative analysis of the performance of these two approaches is presented. We then conduct a failure analysis on selected features to try and understand why the gender classification system failed to predict the tweet tags correctly.

6.7.1 Performance Results Analysis

Feature-engineering Results

Table 6.22 provides a comparison of accuracy scores for different feature sets using various classifiers on both unprocessed (UP) and preprocessed (PP) text data. In the gender classification task, emoji unigrams, TFIDF, and word embeddings each achieved the highest scores when tested individually. However, when combined, emoji unigrams, TFIDF, and

word embeddings did not achieve the best result. Additionally, when TFIDF was used in combination with the top 7 features selected using mutual information, accuracy improved, while the addition of word embeddings (WE) to the top 7 features had no impact. The best-performing combination, observed with UP text, involved the top 7 features chosen through feature selection using mutual information, along with TFIDF, and the use of Support Vector Machines (SVM). This highlights the important role of feature selection that helped identify the smallest set of the most effective features in our dataset that achieved the highest accuracy score in our gender classification task.

However, before we can confidently say that this is the best-performing model, it is important to ensure that there is a statistically significant difference in the performance of the models. To this end, a statistical analysis was carried out using the Mann-Whitney Test to determine if there is a significant difference between the performance of the models. We performed a stratified 10-fold cross-validation for each model and recorded the accuracy score for each fold using the SVM classifier with the UP data. These results were then compared pairwise. Details of the statistical analysis are provided below.

Statistical Analysis The Wilcoxon signed-rank test in table 6.24 revealed several significant differences between the feature sets. We consider p-values two p-values (0.01 and 0.5) as significant thresholds. The combinations that showed significant differences when compared are:

- **Top 7 + TFIDF** significantly outperformed both **Top 7 + WE** and **TF-IDF + emoji unigrams + WE**.
- **Top 7 + TFIDF + WE** also significantly outperformed both **Top 7 + WE** and **TF-IDF + emoji unigrams + WE**.
- **Top 7 + WE** significantly outperformed **TF-IDF + emoji unigrams + WE**.
- **All Features** significantly outperformed both **TF-IDF + emoji unigrams + WE** and **Top 7 + WE**.

However, the combinations that had no significant differences between them are:

- **Top 7 + TFIDF** and **Top 7 + TFIDF + WE**.
- **Top 7 + TFIDF** and **All Features**.
- **Top 7 + TFIDF + WE** and **All Features**.

Overall, the features **Top 7 + TFIDF**, **Top 7 + TFIDF + WE**, and **All Features** stand out as the best performing feature combinations. However, since we want the smallest feature size that achieves the highest accuracy, we consider **Top7 + TFIDF** as the best feature combination for the gender classification task.

	Top 7 + TFIDF	Top 7 + TFIDF + WE	Top 7 + WE	TF-IDF + emoji unigrams + WE
Top 7 + TFIDF	-	0.7695	0.0039*	0.0020**
Top 7 + TFIDF + WE	0.7695	-	0.0039*	0.0020**
Top 7 + WE	0.0039*	0.0039*	-	0.0039*
TF-IDF + emoji unigrams + WE	0.0020**	0.0020**	0.0039*	-
All Features	0.3750	0.3223	0.0020**	0.0020**

Table 6.24 Pairwise p-values from Wilcoxon Signed-Rank Test for Different Feature Sets. * $0.01 < P \leq 0.05$, ** $P \leq 0.01$

Deep Learning Results

Table 6.23 presents a comparison of the transformer models performance in gender classification. Overall, the transformer models show varying performance in different datasets, with Marbert achieving the highest accuracy in the unprocessed (UP) dataset (0.73). Following closely is CAMELBERT, which also performed well on the (UP) dataset (0.72 accuracy). QARiB achieved the lowest accuracy score when used with the (UP) dataset (0.66). One reason why Marbert performed better than the other Arabic transformer models may be because it was trained on 1 billion tweets from Twitter and these tweets contained tweets that have non-Arabic words. During the data collection process, the researchers kept any tweet that had non-arabic words as long as the tweet included a minimum of 3 Arabic words. This means that code-switching was also captured during the pre-training stage of the model and due to our (UP) dataset containing English words, we assume that Marbert was more efficient than the other models that were trained only on Arabic words. CAMELBERT was the second best performing model on the (UP) dataset and one feature that distinguishes it from the other two models is that it has been pretrained on diverse types of dialectal corpora including transcripts, online commentaries, tweets and other dialectal corpora.

Experimenting with two different approaches to build our gender classification system yielded different results and insights. Using a feature engineering approach, the highest accuracy score we achieved was 0.69 using TFIDF in combination with the top 7 features which are the combination of sentiment lexicon features, characteristic vocabulary, code-switching, vocabulary and sentence length, parts of speech, emoji unigrams, and punctuation counts. However, employing a deep learning approach using pre-trained Arabic Transformer models led to even higher accuracy, reaching 0.73. This was accomplished by fine-tuning

the Marbert model with learning rate of $7.658062e - 06$, trained for 3 epochs and maximum sequence length set to 128 tokens. Notably, the deep learning approach using Marbert outperformed the feature engineering approach in our experimentation, likely due to the model's pre-training on a large Twitter dataset of dialectal Arabic. This extensive pre-training, involving the random sampling of one billion tweets from a total of 15.6 billion tweets, may have helped in enabling Marbert to capture some characteristics of dialectal Kuwaiti Arabic as used on Twitter.

6.7.2 Failure Analysis

In order to evaluate our developed gender classification systems, confusion matrices were generated to allow further analysis. Confusion matrices provide information regarding True positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN) (Maria Navin and Pankaja, 2016). In our context, we are analysing the performance of the SVM classifier used for different feature combinations as it was the classifier that achieved the highest accuracy score amongst the other classifiers. In the following analysis:

- (TP) represents the number of female tweets that were correctly predicted as female tweets.
- (TN) represents the number of male tweets that were predicted as male tweets.
- (FP) represents the number of female tweets that were missclassified as male tweets.
- (FN) represents the number of male tweets that were missclassified as female tweets.

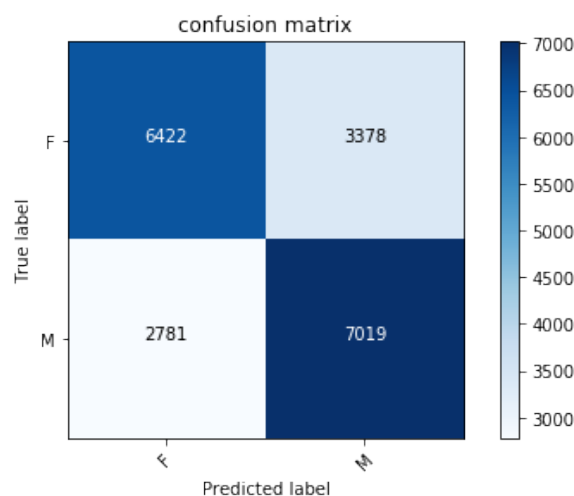


Figure 6.6 Confusion Matrix for Top 7 Features and TFIDF on (UP) Tweets

As the top 7 engineered features in combination with TFIDF was one of the feature combinations that achieved the highest accuracy score of 0.69 on unprocessed tweets, we generated a confusion matrix for this feature combination to analyse the SVM classifier's performance and discuss reasons that may have affected the classifier's performance.

As can be seen in Fig. 6.6, the model's performance varies between the classification of female and male tweets. In the classification of female tweets, the model correctly predicted 6422 tweets as female (TP), but misclassified 3378 tweets as male (FN). This indicates that while the model is relatively successful in identifying female tweets, there is a notable number of female tweets that are being incorrectly classified as male. Conversely, in the classification of male tweets, the model correctly predicted 7019 tweets as male (TN), but misclassified 2781 tweets as female (FP). This suggests that the model is generally effective in identifying male tweets, although there are still a significant number of misclassifications in comparison to the female class.

We selected the sentiment lexicon feature and the characteristic vocabulary feature for failure analysis because they were among the top 7 features that contributed to the performance of our gender classification model. These features hold particular linguistic interest, as they are key to capturing emotional expression and gender-specific language use.

- **Sentiment Lexicon Feature:** For the lexicon based variation of this feature implementation, we investigated the words in our dataset that were not found in the lexicons we used, we did this on both the pre-processed and unprocessed text. For the pre-processed text, 17002 words out of 64321 were not found in the lexicon. We also tried to see what types of words that were not found, we found that English words, some clitics (the output of the pre-processing step) and some new vocabulary that have been trending in social media in the past few years such as *vaccines*.

For the Unprocessed text, 50324 words out of 185605 were not found in the lexicon. These words include emojis, words concatenated with punctuation, @USER and URL special tokens that were added during pre-processing the tweets to anonymise the users, diacritised words, lengthened words.

- **Characteristic Vocabulary Feature:**

We conducted an error analysis on the characteristic vocabulary feature to assess how well it was performing in distinguishing between male and female tweets. Specifically, we examined words that were not assigned to either the male or female vocabulary lists to determine their quantity and potential impact on the model's performance. We found that 34,324 words were not included in any of the gender-specific lists, which could lead to misclassifications. These unassigned words may include gender-neutral

terms, rare or emerging vocabulary, or words that are equally used by both genders. We noticed that 34324 words were not assigned to any of the gender lists. Figure 6.7 shows the confusion matrix generated using this feature with SVM classifier. The model seems to perform better at identifying tweets by males (6857 from 9800) than from females (5135 from 9800).

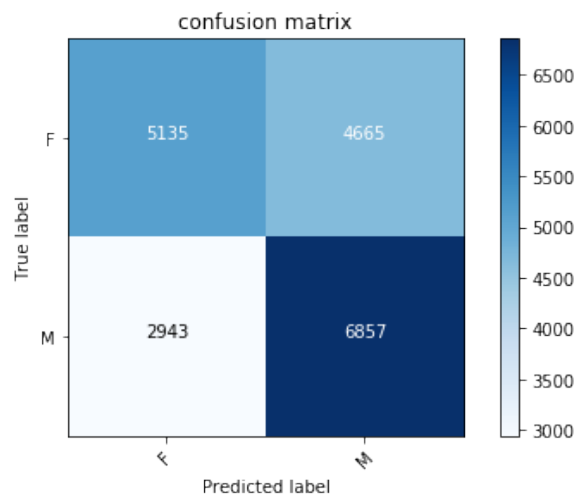


Figure 6.7 Confusion Matrix for Characteristic Vocabulary on (UP) Tweets

6.8 Summary

In summary, this chapter presented the compilation of the Kuwaiti Arabic Twitter Dataset (KATD): the first and largest publicly available gender-labelled dataset of Kuwaiti Arabic tweets. The data collection and labelling processes were explained in this chapter. Then, two supervised learning approaches to building a gender classification system: a feature engineering approach and a deep learning approach were discussed. The feature engineering approach involved experimenting with different statistical and linguistic features using machine learning algorithms and achieved an accuracy score of 0.69 using TFIDF and 7 top selected features with an SVM classifier on unprocessed text. The deep learning approach involved fine-tuning pre-trained Arabic Transformer models and the model that achieved the highest accuracy score was Marbert (0.73) accuracy on unprocessed text. We concluded this chapter with a failure analysis to find explanations as to why certain features did not perform well.

There are currently no existing studies on gender classification specifically for Kuwaiti Arabic to directly compare our results to. However, in general, gender classification studies on English texts have reported accuracy ranges from 0.72 to 0.80, such as Rao et al. (2010)

and Suero Montero et al. (2014), respectively. As for Arabic gender classification systems, reported accuracies range from 0.65 in Abdul-Mageed et al. (2019) for Arabic tweets to 0.93 in Al-Ghadir and Azmi (2019) for Saudi social media posts. Although our results are slightly below the highest reported accuracies, it is important to consider that Al-Ghadir and Azmi's study was conducted on online posts, which generally contain more information per instance compared to tweets. This means that their classifier had access to a richer set of linguistic features to learn from, which contributes to their higher accuracy. In contrast, our classifier faced the unique linguistic challenges of Kuwaiti Arabic within the more constrained format of tweets.

Chapter 7

Conclusion and Future Work

In this thesis, a computational linguistic approach has been undertaken to study the relationship between language and gender in the Kuwaiti Arabic dialect on online social media platforms. In this chapter, we review the key contributions of the thesis and revisit the research questions that guided the studies conducted. We also outline suggestions for potential future work in this growing field of study.

7.1 Summary of Contributions

This study makes a significant contribution to the field of Arabic Natural Language Processing and offers valuable insights for scholars exploring sociolinguistic dimensions such as gender and its interplay with language use. The contributions of this study can be categorised into three primary areas: the compilation of unique, publicly available datasets for the Kuwaiti Arabic dialect; the development of state-of-the-art gender classification systems using KA textual datasets; and, the drawing of insights into the interplay between language and gender within this specific cultural and digital context.

7.1.1 Kuwaiti Arabic Datasets

This thesis addresses the paucity of Kuwaiti Arabic datasets that are publicly available for research purposes. We have compiled and annotated three social media datasets of Kuwaiti Arabic that have been made publicly available ¹ for researchers in the field. Each dataset is unique in its characteristics and what it offers to the research field:

- KAGen: is the first publicly available dataset of conversational Kuwaiti Arabic compiled from WhatsApp reading club groups consisting of mixed gender Kuwaiti users.

¹The datasets will be published through the University of Sheffield's research data repository (ORDA)

It has been labelled with gender information and is a valuable resource for researchers interested in studying gender dynamics and conversation analysis in KA. KAGen is described in detail in section 4.3.1.

- **KACD:** is an extension of KAGen that enriches the dataset with additional annotations. It has been annotated following a developed framework for annotating conversational data. KACD is the first publicly available dataset of conversational KA annotated with conversational function tags. KACD is described in detail in section 5.3.4.
- **KATD:** Is the largest and first publicly available dataset of KA tweets labelled by gender. It serves as a foundational resource for researchers investigating gender-related linguistic phenomena and sociolinguistic dynamics within Kuwaiti Arabic discourse. Additionally, KATD can significantly aid NLP researchers aiming to fine-tune and evaluate large language models on Kuwaiti Arabic text especially given that dialectal Kuwaiti Arabic data is not abundantly available for training large language models (LLMs). KATD is described in detail in section 6.3.

7.1.2 Gender Classification Systems

Two gender classification systems were developed using two of our compiled datasets (KAGen and KATD). We experimented with two supervised learning approaches: (a) a feature-engineering approach guided by sociolinguistic and statistical features elicited from the literature and research field, and (b) a deep learning approach using pre-trained Arabic transformer models that have been fine-tuned to our task.

7.1.3 Insights into Gendered Linguistic Interactions in KA

The thesis unfolds in three main studies, each of which provides insights for our research questions.

RQ1: *What distinguishes the language use of Kuwaiti female and male social media users?*

This question was addressed by the creation of the KAGen and the KATD datasets. In KAGen, we compiled a dataset of conversational KA extracted from Kuwaiti WhatsApp reading club groups of mixed gender users. We performed a qualitative and quantitative analysis of the dataset (refer to sec. 4.4) to study if there are differences in certain features related to the language of Kuwaiti men and women social media users. We looked into various interactional features such as average utterance length, number of turns, average emoji usage, and gender exclusive words. Our findings show that men and women differ in

their interactional patterns pertaining to number of turns taken and emoji usage where men showed higher activity levels in turn-taking while women demonstrated higher emoji use. Our qualitative analysis also drew interesting insights, including the linguistic phenomenon of elongation or use of stretched words where women were noticed to use these linguistic devices more than men and in a wider range of linguistic settings. However, men used stretched words less than women and in certain contexts when greeting and laughing. Distinct patterns of emoji use were also observed including the types of emojis used by each gender and certain emoji combinations were used with different frequencies amongst men and women.

In KATD, a large scale dataset of KA tweets was compiled and labelled with gender labels. This dataset was then explored for gender characteristic vocabulary and emojis. We used the chi-square test to identify lists of gender associated vocabulary and emojis and observed distinct patterns. Terms related to sports, politics, economics, group camaraderie were found to be characteristic of men's language. While women's characteristic vocabulary revolved around social greeting terms, emotional terms, religious terms, and beauty related terms. Comparing the usage patterns, female-associated emojis were predominantly centered around themes of love, affection, and emotional expression, with a significant presence of hearts and other emotive symbols. This suggests a communication style that values connectivity and emotional transparency. On the other hand, male-associated emojis were related to humor, approval, and action, with a strong presence of laughter, thumbs up, and physical activities. This indicates a preference for lighter, more affirming, and active interactions.

RQ2: *Are there specific conversational strategies employed by Kuwaiti male and female users in WhatsApp exchanges and do they vary between both gender groups?*

This question was addressed by creating (KACD) in which we developed a framework to annotate conversational Kuwaiti Arabic used in WhatsApp reading club groups of mixed gender Kuwaiti users. We conducted a thematic analysis to derive the conversational functions used in the conversations and our analysis resulted with seven most prominent conversational functions: Arranging Club Meeting, Book Discussion, General-Reading-related Discussion, Feedback on Club Meeting, Social Interaction, Greeting and Leave-taking. We conducted a statistical analysis to study if there are differences regarding the proportions of utterances assigned to each conversational function tag between men and women. Results showed that there is no strong evidence of a difference in the proportion of utterances between men and women in the conversational functions we studied. However, there were indications of divergence in the usage of Feedback on Club Meeting and Book Discussion, with men exhibiting a higher proportion of conversational function tags in these categories, possibly due to a proactive approach to engagement.

Additionally, our qualitative analysis of the language used under the Feedback on Club Meeting tag revealed notable distinctions in the language usage and emoji usage between men and women. We noticed that men tended to use Modern Standard Arabic and fewer emojis, while women used Kuwaiti Arabic and employed emojis, particularly heart emojis, thumbs up emojis, and smileys, more frequently.

RQ3: *To what extent can the gender of Kuwaiti social media users be predicted from their online language use?*

This question was addressed in two of our studies by developing gender classification systems trained and evaluated on our compiled datasets (KAGen and KATD). In KAGen, we developed a basic gender classification system using a supervised feature engineering approach. The features experimented with in this system were derived from our quantitative and qualitative analysis of the KAGen dataset. We experimented with various linguistic and statistical features such as emoji bigrams, word counts, stretched words, and length of turns. We performed 10-fold cross-validation on the dataset and achieved the highest balanced accuracy score of 0.67 when using a combination of word counts, turn length, stretched words and emoji bigrams on unprocessed text.

In KATD, we experimented with building a gender classification system using a feature engineering approach and a deep-learning approach using pre-trained Arabic Transformer models that were fine-tuned to our task. In our feature engineering approach, various feature sets were evaluated using traditional text classification classifiers (SVMs, Logistic Regression, KNN, and Decision Trees). Results showed that combining TFIDF with the top 7 features selected through mutual information (sentiment lexicon features, characteristic vocabulary, code-switching, vocabulary and sentence length, parts of speech, emoji unigrams, and punctuation counts) achieved the highest accuracy score of 0.69 on unprocessed text with an SVM classifier (notably higher than the random choice baseline figure of 0.50, given that the dataset contains equal numbers of male and female tweets). Moreover, a statistical analysis was conducted to examine whether significant differences existed among the various feature sets experimented with and results showed that the combination TFIDF with the top 7 features was the most significant combination compared to all other combinations.

In our deep learning approach, we experimented with fine-tuning different transformer models pre-trained on dialectal Arabic such as Marbert, CAMeLBERT, and QARiB. The model that achieved the highest accuracy score was Marbert which achieved an accuracy score of 0.73 when using a learning rate of $7.658062e - 06$, trained for 3 epochs and maximum sequence length set to 128 tokens.

We can conclude that the models developed using two different supervised learning approaches on our large scale dataset (KATD): the feature engineering approach and the

deep learning approach showed close performance in predicting the gender of Kuwaiti social media users based on their online language use. However, the superior performance of the deep learning approach can be credited to its capability to capture non-obvious linguistic patterns and differences within the data.

7.2 Future Work

There are numerous potential avenues for future research that extend beyond the scope of the current study. These explorations could provide further insights and advancements in the field of gender classification in Kuwaiti Arabic dialects on social media. Some of these include:

- One future work suggestion would be to expand KACD by gathering more data through recording and transcribing face-to-face book club meetings. We can then compare text-based and spoken interactions. Additionally, if we expand the dataset we can perform further investigation into the weak correlations found between two of the tags and male usage, potentially uncovering deeper insights.
- Furthermore, the findings of this thesis demonstrate promising effects of using gender-specific vocabulary and emojis as features to train gender classification systems for the Kuwaiti Arabic dialect. For future work, expanding the creation of Kuwaiti Arabic lexicons by incorporating a broader range of gender-specific vocabulary and emojis could further enhance these systems. This can be achieved by extracting these linguistic devices from various sources such as social media data, user surveys, and online forums. Such an approach may improve the performance and accuracy of gender classification systems.
- An interesting future experiment could involve establishing a human upper bound on performance in the gender classification task. This would involve analysing how well humans perform in identifying gender based on language use to compare the effectiveness of computational models. Such an experiment would offer valuable insights into the strengths and limitations of both approaches.
- It would also be interesting to explore the characteristics of misclassified tweets, such as their length compared to correctly classified ones to see if misclassified tweets tend to be shorter on average, which might indicate specific challenges for gender classification algorithms. Performing this type of failure analysis may help improve the

accuracy of these algorithms by understanding how tweet length impacts classification performance.

- One potential future work direction could be considering an alternative approach to gender classification which involves using user-level classification rather than tweet-level classification. A simple approach to this would be to classify all of a user's tweets individually and then assign the majority class to the user. One could then reclassify each tweet with the user level gender tag and see if this approach outperforms the tweet-based classification.
- Another potential contribution for future work involves addressing the current lack of sufficient Kuwaiti Arabic text in dialectal datasets used to pre-train large language models (LLMs). Collecting more Kuwaiti Arabic textual data could significantly enhance the performance of these models when processing this specific dialect. By providing KA datasets, the effectiveness and accuracy of LLMs in understanding and processing Kuwaiti Arabic can be improved.

7.3 Concluding Remarks

This thesis has contributed to the fascinating and rapidly evolving field of computational linguistics, particularly focusing on gender analysis and classification of the Kuwaiti Arabic dialect on social media platforms. We carefully compiled datasets of the Kuwaiti Arabic dialect and conducted several experiments that examined linguistic gender differences between Kuwaiti men and women. We employed a variety of methods: quantitative, qualitative, and state-of-the-art computational tools to study and analyse language and gender dynamics in the use of Kuwaiti Arabic in online social media platforms. We experimented with several developed machine learning models to perform the task of predicting the gender of Kuwaiti social media users based on their language use. Through these experiments, we gained valuable insights into the specific linguistic differences and communicative patterns of interaction between genders in this dialect.

In summary, this thesis represents a pioneering effort to explore Kuwaiti Arabic and its linguistic gender dynamics through a blend of computational linguistics techniques and sociolinguistic insights. Additionally, it lays the groundwork for future advancements in the research field by offering new avenues to explore and refine gender classification tasks. Moreover, it aims to provide valuable resources for scholars and researchers engaged in language and gender studies, thus contributing to the broader discourse on this fascinating subject.

Bibliography

- Ahmed Abdelali. Localization in modern standard arabic. *Journal of the American Society for Information Science and technology*, 55(1):23–28, 2004.
- Ahmed Abdelali, James Cowie, and Hamdy Soliman. Building a modern standard arabic corpus. In *Workshop on computational modeling of lexical acquisition*, pages 25–28, 2005.
- Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, 2016.
- Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. Pre-training bert on arabic tweets: Practical considerations. 2021.
- Muhammad Abdul-Mageed, Chiyu Zhang, Arun Rajendran, AbdelRahim Elmadany, Michael Przysupka, and Lyle Ungar. Sentence-level bert and multi-task learning of age and gender in social media. *arXiv preprint arXiv:1911.00637*, 2019.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.551. URL <https://aclanthology.org/2021.acl-long.551>.
- Ghazi Abuhakema, Reem Faraj, Anna Feldman, and Eileen Fitzpatrick. Annotating an arabic learner corpus for error. 2008.
- Shaimaa Ben Aichaoui, Nawel Hiri, Abdelhalim Hafedh Dahou, and Mohamed Amine Cheragui. Automatic building of a large arabic spelling error corpus. *SN Computer Science*, 4(2):108, 2022.
- Abdulrahman I Al-Ghadir and Aqil M Azmi. A study of arabic social media users—posting behavior and author’s gender prediction. *Cognitive Computation*, 11:71–86, 2019.
- Shamlan Al-Qenaie et al. *Kuwaiti Arabic: A socio-phonological perspective*. PhD thesis, Durham University, 2011.
- Latifa Al-Sulaiti and Eric Steven Atwell. The design of a corpus of contemporary arabic. *International journal of corpus linguistics*, 11(2):135–171, 2006.

- Latifa Al-Sulaiti and ES Atwell. Extending the corpus of contemporary arabic. In *Proceedings of the CL'2005 Corpus Linguistics Conference*. UCREL, Lancaster University, 2005.
- Enam Al-Wer. Language and gender in the middle east and north africa. 2014.
- Sameh Alansary and Magdi Nagi. The international corpus of arabic: Compilation, analysis and evaluation. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 8–17, 2014.
- AYG Alfaifi, Eric Atwell, and Ibraheem Hedaya. Arabic learner corpus (alc) v2: a new written and spoken corpus of arabic learners. In *Proceedings of Learner Corpus Studies in Asia and the World 2014*, volume 2, pages 77–89. Kobe International Communication Center, 2014.
- Nada Algharabali. *Two cultures, one room: investigating language and gender in Kuwait*. PhD thesis, 2010.
- Lafi M Alharbi. Formal analysis of intonation: The case of the kuwaiti dialect of arabic. 1992.
- Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.
- Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. Dart: A large dataset of dialectal arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- Kholoud Alsmearat, Mahmoud Al-Ayyoub, and Riyad Al-Shalabi. An extensive study of the bag-of-words approach for gender identification of arabic articles. In *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, pages 601–608. IEEE, 2014.
- Jo Angouri and Judith Baxter. *The routledge handbook of language, gender, and sexuality*. Routledge, 2021.
- Tharwat Arafat and Bilal Hamamra. Gender and word elongation in facebook-mediated communication in palestinian arabic. *Communication Research and Practice*, 7(3):221–242, 2021.
- Christina Aravantinou, Vasiliki Simaki, Iosif Mporas, and Vasileios Megalooikonomou. Gender classification of web authors using feature selection and language models. In *Speech and Computer: 17th International Conference, SPECOM 2015, Athens, Greece, September 20-24, 2015, Proceedings 17*, pages 226–233. Springer, 2015.
- David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.
- Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*, 2017.

- Reem Bassiouney. *Arabic sociolinguistics: Topics in diglossia, gender, identity, and politics*. Georgetown University Press, 2020.
- Yonatan Belinkov, Alexander Magidow, Maxim Romanov, Avi Shmidman, and Moshe Koppel. Shamela: A large-scale historical arabic corpus. *arXiv preprint arXiv:1612.08989*, 2016.
- Fadi Biadisy, Julia Hirschberg, and Nizar Habash. Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, pages 53–61. Association for Computational Linguistics, 2009.
- Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander, and Owen Rambow. Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP)*, pages 93–103, 2014.
- Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O’Reilly Media Inc., 2009.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghrouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.
- Naaima Boudad, Rdouan Faizi, Rachid Oulad Haj Thami, and Raddouane Chiheb. Sentiment analysis in arabic: A review of the literature. *Ain Shams Engineering Journal*, 9(4): 2479–2490, 2018.
- Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.
- Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.
- John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1301–1309, 2011.
- Deborah Cameron. Gender, language, and discourse: A review essay. *Signs: Journal of Women in culture and society*, 23(4):945–973, 1998.
- Colin Campbell and Yiming Ying. *Learning with support vector machines*. Springer Nature, 2022.
- Bahzad Charbuty and Adnan Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01):20–28, 2021.

- Zhenpeng Chen, Xuan Lu, Sheng Shen, Wei Ai, Xuanzhe Liu, and Qiaozhu Mei. Through a gender lens: An empirical study of emoji usage over large-scale android users. *arXiv preprint arXiv:1705.05546*, 2017.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- Mita K Dalal and Mukesh A Zaveri. Automatic text classification: a technical review. *International journal of computer applications*, 28(2):37–40, 2011.
- Ali A Dashti, Hamed H Al-Abdullah, and Hasan A Johar. Social media and the spiral of silence: The case of kuwaiti female students political discourse on twitter. *Journal of International Women's Studies*, 16(3):42–53, 2015.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018a.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018b.
- Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. Colaba: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing*, pages 66–74. Citeseer, 2010.
- Penelope Eckert and Sally McConnell-Ginet. *Language and gender*. Cambridge University Press, 2013.
- Samhaa R El-Beltagy. Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2900–2905, 2016.
- AbdelRahim Elmadany, Hamdy Mubarak, and Walid Magdy. Arsas: An arabic speech-act and sentiment corpus of tweets. *OSACT*, 3:20, 2018.
- Ashraf Elnagar, Sane M Yagi, Ali Bou Nassif, Ismail Shahin, and Said A Salloum. Systematic literature review of dialectal arabic: identification and detection. *IEEE Access*, 9:31010–31042, 2021.
- Nadir Omer Fadl Elssied, Othman Ibrahim, and Ahmed Hamza Osman. A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7(3):625–638, 2014.
- Ali Farghaly and Khaled Shaalan. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4): 1–22, 2009.
- Ibrahim Abu Farha and Walid Magdy. Mazajak: An online arabic sentiment analyser. In *Proceedings of the fourth arabic natural language processing workshop*, pages 192–198, 2019.

- Wolfdietrich Fischer. 11 classical arabic. *The Semitic languages*, page 187, 2013.
- Karèn Fort. *Collaborative annotation for reliable natural language processing: Technical and sociological aspects*. John Wiley & Sons, 2016.
- Eduard Fosch-Villaronga, Adam Poulsen, Roger Andre Søråa, and BHM Custers. A little bird told me your gender: Gender inferences in social media. *Information Processing & Management*, 58(3):102541, 2021.
- Mohammed M Fouad, Ahmed Mahany, Naif Aljohani, Rabeeh Ayaz Abbasi, and Saeed-Ul Hassan. Arwordvec: efficient word embedding models for arabic tweets. *Soft Computing*, 24:8061–8068, 2020.
- Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83, 2022.
- Aurélien Geron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.", 2022.
- Shazia Akbar Ghilzai and Mahvish Baloch. Conversational analysis of turn taking behavior and gender differences in multimodal conversation. *European Academic Research*, 3(9): 10100–10116, 2015.
- Evelyn Gius, Jan Christoph Meister, Malte Meister, Marco Petris, Mareike Schumacher, and Dominik Gerstorfer. Catma, May 2023. URL <https://doi.org/10.5281/zenodo.7986177>.
- Orly Turgeman Goldshmidt and Leonard Weller. “talking emotions”: Gender differences in a variety of conversational contexts. *Symbolic Interaction*, 23(2):117–134, 2000.
- Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507, 2021.
- Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. Guidelines for annotation of arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53, 2008.
- Nizar Y Habash. Introduction to arabic natural language processing. *Synthesis lectures on human language technologies*, 3(1):1–187, 2010.
- Yaakov HaCohen-Kerner. Survey on profiling age and gender of text authors. *Expert Systems with Applications*, 199:117140, 2022.
- Emma Haddi, Xiaohui Liu, and Yong Shi. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26 – 32, 2013. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2013.05.005>. URL <http://www.sciencedirect.com/science/article/pii/S1877050913001385>. First International Conference on Information Technology and Quantitative Management.
- Allan Hanbury, Andreas Rauber, and Arjen Vries. *Multidisciplinary Information Retrieval: Second Information Retrieval Facility Conference, IRFC 2011, Vienna, Austria, June 6, 2011, Proceedings*, volume 6653. Springer Science & Business Media, 2011.

- Sayar Ul Hassan, Jameel Ahamed, and Khaleel Ahmad. Analytics of machine learning-based algorithms for text classification. *Sustainable operations and computers*, 3:238–248, 2022.
- Janet Holmes and Miriam Meyerhoff. Different voices, different views: An introduction to current research in language and gender. *The handbook of language and gender*, pages 1–17, 2003.
- Janet Holmes and Miriam Meyerhoff. *The handbook of language and gender*, volume 25. John Wiley & Sons, 2008.
- Fatemah Husain, Hana Al-Ostad, and Halima Omar. A weak supervised transfer learning approach for sentiment analysis to the kuwaiti dialect. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 161–173, 2022.
- Shereen Hussein, Mona Farouk, and ElSayed Hemayed. Gender identification of egyptian dialect in twitter. *Egyptian Informatics Journal*, 20(2):109–116, 2019.
- Sadegh Bafandeh Imandoust, Mohammad Bolandraftar, et al. Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International journal of engineering research and applications*, 3(5):605–610, 2013.
- Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*, 2021a.
- Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*, 2021b.
- Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. Curras: an annotated corpus for the palestinian arabic dialect. *Language Resources and Evaluation*, 51:745–775, 2017.
- Rajni Jindal, Ruchika Malhotra, and Abha Jain. Techniques for text classification: Literature review and current trends. *webology*, 12(2), 2015.
- Kyle P. Johnson, Patrick Burns, John Stewart, and Todd Cook. Cltk: The classical language toolkit, 2014–2020. URL <https://github.com/cltk/cltk>.
- Daniel Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, January 7 2023. URL <https://web.stanford.edu/~jurafsky/slp3/>. Bug-fixing and restructuring release.
- Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. A large scale corpus of gulf arabic. *arXiv preprint arXiv:1609.02960*, 2016.
- Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. A morphologically annotated corpus of emirati arabic. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

- Salam Khalifa, Nasser Zalmout, and Nizar Habash. Morphological analysis and disambiguation for gulf arabic: The interplay between resources and methods. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3895–3904, 2020.
- Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002.
- Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.
- Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. Shami: A corpus of levantine arabic dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.
- Robin Lakoff. Language and woman’s place. *Language in society*, 2(1):45–79, 1973.
- Lia Litosseliti and Jane Sunderland. *Gender identity and discourse analysis*, volume 2. John Benjamins Publishing, 2002.
- Jan Luts, Fabian Ojeda, Raf Van de Plas, Bart De Moor, Sabine Van Huffel, and Johan AK Suykens. A tutorial on support vector machine-based methods for classification problems in chemometrics. *Analytica chimica acta*, 665(2):129–145, 2010.
- Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona T Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. Developing and using a pilot dialectal arabic treebank. In *LREC*, pages 443–448, 2006.
- Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development. In *LREC*, pages 2348–2354, 2014.
- Yasser Mahgoub. Globalization and the built environment in kuwait. *Habitat International*, 28(4):505–519, 2004.
- Arun S. Maiya. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*, 2020.
- Arun S Maiya. ktrain: A low-code library for augmented machine learning. *The Journal of Machine Learning Research*, 23(1):7070–7075, 2022.
- Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.
- Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
- JR Maria Navin and R Pankaja. Performance analysis of text classification algorithms using confusion matrix. *International Journal of Engineering and Technical Research (IJETR)*, 6(4):75–8, 2016.

- Stephen Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2011.
- Viera Maslej-Krešňáková, Martin Sarnovský, Peter Butka, and Kristína Machová. Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. *Applied Sciences*, 10(23):8631, 2020.
- Peter Hugoe Matthews. *Morphology*. Cambridge Univ. Press, 2009.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- Saif Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. Sentiment lexicons for arabic social media. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, pages 33–37, 2016.
- Aiman Moldagulova and Rosnafisah Bte Sulaiman. Using knn algorithm for classification of textual documents. In *2017 8th international conference on information technology (ICIT)*, pages 665–671. IEEE, 2017.
- Hamdy Mubarak, Shammur Absar Chowdhury, and Firoj Alam. Arabgend: Gender analysis and inference on arabic twitter. *arXiv preprint arXiv:2203.00271*, 2022.
- Mariana Neves and Jurica Ševa. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1):146–163, 2021.
- Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.
- Elisabeth Noelle-Neumann. *The spiral of silence: Public opinion, our social skin*. University of Chicago Press, 1993.
- Michael Oakes, Robert Gaaizauskas, Helene Fowkes, Anna Jonsson, Vincent Wan, and Micheline Beaulieu. A method based on the chi-square test for document classification. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 440–441, 2001.
- Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032, 2020.
- Sarra Ouni, Fethi Fkih, and Mohamed Nazih Omri. A survey of machine learning-based author profiling from texts analysis in social networks. *Multimedia Tools and Applications*, pages 1–34, 2023.
- Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf.

- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.
- Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, pages 1613–0073, 2017.
- Francisco Rangel, Paolo Rosso, Anis Charfi, Wajdi Zaghrouani, Bilal Ghanem, and Javier Sánchez-Junquera. Overview of the track on author profiling and deception detection in arabic. *Working Notes of FIRE 2019. CEUR-WS. org, vol. 2517*, pages 70–83, 2019.
- Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44, 2010.
- Paul Rayson, Geoffrey N Leech, and Mary Hodges. Social differentiation in the use of english vocabulary: some analyses of the conversational component of the british national corpus. *International Journal of Corpus Linguistics*, 2(1):133–152, 1997.
- Eshrag Refaee and Verena Rieser. An arabic twitter corpus for subjectivity and sentiment analysis. In *9th International Language Resources and Evaluation Conference*, pages 2268–2273. European Language Resources Association, 2014.
- Simon Rogers and Mark Girolami. *A first course in machine learning*. CRC Press, 2016.
- Avi Rosenfeld, Sigal Sina, David Sarne, Or Avidov, and Sarit Kraus. Whatsapp usage patterns and prediction models. In *ICWSM/IUSSP Workshop on Social Media and Demographic Research*, 2016.
- Motaz K Saad and Wesam Ashour. Osac: Open source arabic corpora. In *6th ArchEng Int. Symposiums, EEECS*, volume 10, page 55, 2010.
- Mohammad Salameh Saif M. Mohammad and Svetlana Kiritchenko. Sentiment lexicons for arabic social media. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia, 2016.
- Jana Ben Salamah and Aymen Elkhlifi. Microblogging opinion mining approach for kuwaiti dialect. In *The International Conference on Computing Technology and Information Management (ICCTIM)*, page 388. Citeseer, 2014.
- Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

- Abdul-Baquee M Sharaf and Eric Atwell. Qurana: Corpus of the quran annotated with pronominal anaphora. In *Lrec*, pages 130–137, 2012a.
- Abdul-Baquee M Sharaf and Eric Atwell. Qursim: A corpus for evaluation of relatedness in short texts. In *LREC*, pages 2295–2302, 2012b.
- Roger W Shuy. A brief history of american sociolinguistics, 1949–1989. *Sociolinguistics: The essential readings*, pages 4–16, 2003.
- Calkin Suero Montero, Myriam Munezero, and Tuomo Kakkonen. Investigating the role of emotion-based features in author gender classification of text. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 98–114. Springer, 2014.
- Reem Suwaileh, Mucahid Kutlu, Nihal Fathima, Tamer Elsayed, and Matthew Lease. Arab-icweb16: A new crawl for today’s arabic web. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 673–676, 2016.
- Deborah Tannen. You just don’t understand: Women and men. *Conversation*. New York: Ballantine books, 1990.
- Patrick Adolf Telnoni, Reza Budiawan, and Mutia Qana’a. Comparison of machine learning classification method on text-based case in twitter. In *2019 International Conference on ICT for Smart Society (ICISS)*, volume 7, pages 1–5. IEEE, 2019.
- Rob Thomson, Tamar Murachver, and James Green. Where is the gender in gendered language? *Psychological Science*, 12(2):171–175, 2001.
- Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. *Natural language processing with transformers*. " O’Reilly Media, Inc.", 2022.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24:175–186, 2014.
- Shan Wareing. What do we know about language and gender. In *eleventh sociolinguistic symposium, Cardiff, September*, pages 5–7, 1996.
- Carol Waseleski. Gender and the use of exclamation points in computer-mediated communication: An analysis of exclamations posted to two electronic discussion lists. *Journal of Computer-Mediated Communication*, 11(4):1012–1024, 2006.
- Deborah Wheeler. New media, globalization and kuwaiti national identity. *The Middle East Journal*, pages 432–444, 2000.

- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- Zhang Yun-tao, Gong Ling, and Wang Yong-cheng. An improved tf-idf approach for text classification. *Journal of Zhejiang University-Science A*, 6:49–55, 2005.
- Wajdi Zaghoulani and Anis Charfi. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. *arXiv preprint arXiv:1808.07674*, 2018.
- Wajdi Zaghoulani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. Large scale arabic error annotation: Guidelines and framework. 2014.
- Omar F Zaidan and Chris Callison-Burch. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics, 2011.
- Taha Zerrouki. Tashaphyne, arabic light stemmer/segment, 2010. URL <https://pypi.python.org/pypi/Tashaphyne/0.2>.
- Taha Zerrouki and Amar Balla. Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. *Data in brief*, 11:147–151, 2017.
- Yulei Zhang, Yan Dang, and Hsinchun Chen. Gender classification for web forums. *Ieee Transactions On Systems, Man, And Cybernetics-Part A: Systems And Humans*, 41(4): 668–677, 2011.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

