University of Kerbala
College of Computer Science & Information Technology
Computer Science Department

Gender Classification Based on Linguistic Style Analysis Using Combine Machine Learning and Deep Learning Techniques

A Thesis

Submitted to the Council of the College of Computer Science & Information Technology / University of Kerbala in Partial Fulfillment of the Requirements for the Master Degree in Computer Science

Written by
Haneen Tamim Abd Ali Hashim

Supervised by
Asst. Prof. Dr. Dhamyaa Abbas Habeeb

2024 A.D. 1446 A.H.

بسم الله الرحمن الرحيم

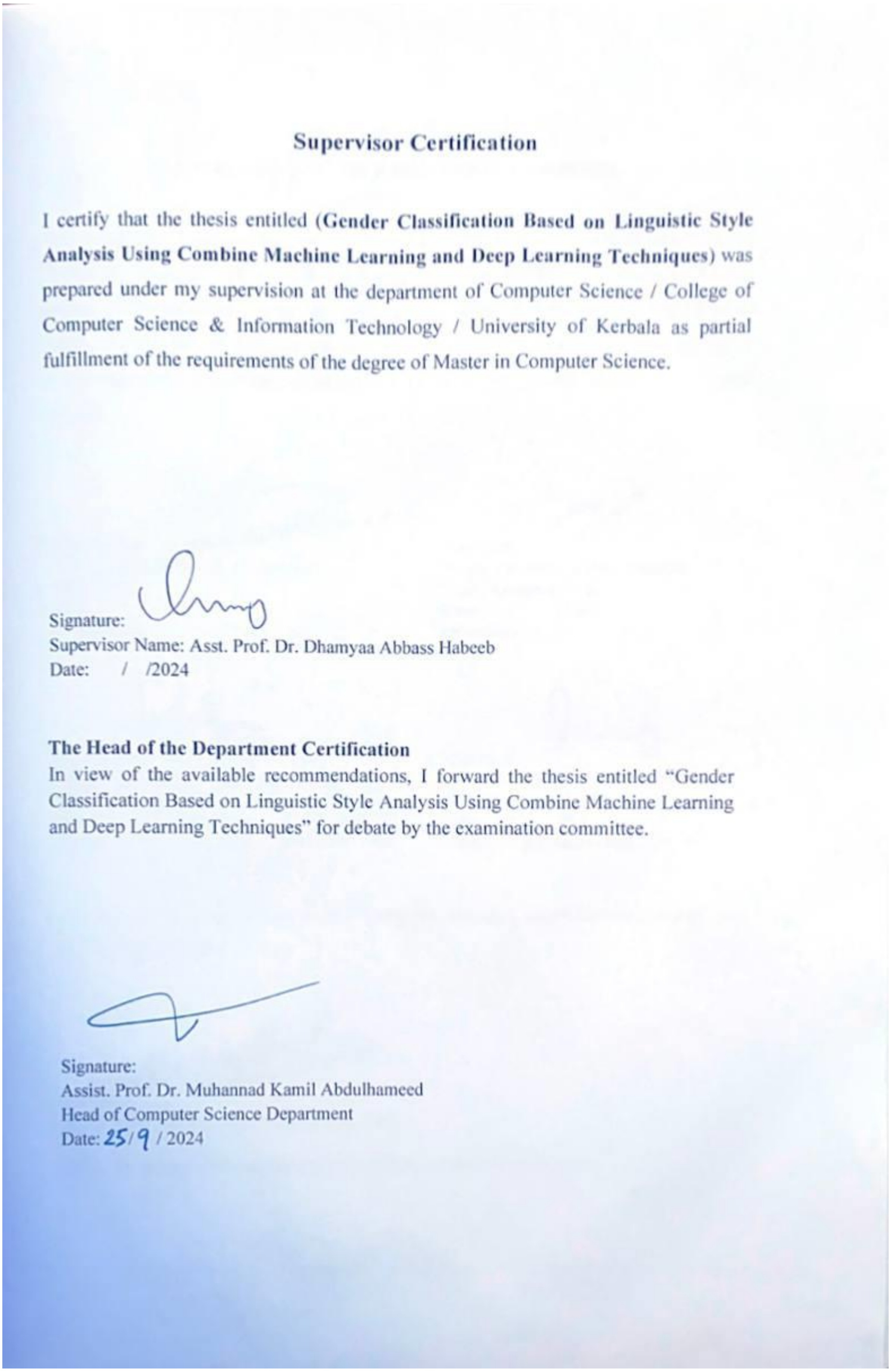
﴿يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ وَاللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ﴾

صدق الله العظيم

سورة المجادلة

آية 11

Supervisor Certification



Supervisor Certification

I certify that the thesis entitled (**Gender Classification Based on Linguistic Style Analysis Using Combine Machine Learning and Deep Learning Techniques**) was prepared under my supervision at the department of Computer Science / College of Computer Science & Information Technology / University of Kerbala as partial fulfillment of the requirements of the degree of Master in Computer Science.

Signature:
Supervisor Name: Asst. Prof. Dr. Dhamyaa Abbass Habeeb
Date: / /2024

The Head of the Department Certification
In view of the available recommendations, I forward the thesis entitled "Gender Classification Based on Linguistic Style Analysis Using Combine Machine Learning and Deep Learning Techniques" for debate by the examination committee.

Signature:
Assist. Prof. Dr. Muhannad Kamil Abdulhameed
Head of Computer Science Department
Date: 25/9 / 2024

Certification of the Examination Committee

We hereby certify that we have studied the dissertation entitled (**Gender Classification Based on Linguistic Style Analysis Using Combine Machine Learning and Deep Learning Techniques**) presented by the student (**Haneen Tamim Abd Ali Hashim**) and examined him/her in its content and what is related to it, and that, in our opinion, it is adequate with (**Excellent**) standing as a thesis for the Master degree in Computer Science.

Signature:
Name: Dr. Noor Dhia Al-Shakarchy
Title: Prof.
Date: / /2024
(Chairman)

Signature:
Name: Dr. Hiba Jabbar Aleqabie
Title: Assistant Prof.
Date: / /2024
(Member)

Signature:
Name: Dr. Akeel Abdulkarim Farhan
Title: Assistant Prof.
Date: / /2024
(Member)

Signature:
Name: Dr. Dhamyaa Abbass Habeeb
Title: Assistant Prof.
Date: / /2024
(Member and Supervisor)

Approved by the Dean of the College of Computer Science & Information Technology, University of Kerbala.

Mow

Signature:
Assist. Prof. Dr. Mowafak Khadom Mohsen
Date: / /2024
(Dean of College of Computer Science & Information Technology)

Dedication

I express deep gratitude to my wonderful family for their immense support and love. Their constant encouragement enabled me to overcome obstacles on my path. I am indebted to my dad, mom, sister, and brother for their invaluable influence on my life. I am forever grateful for their unwavering support in helping me achieve my goals. I hope they will always be proud of me.

Haneen Tamim

Acknowledgement

In the name of Allah, the Most Gracious and Merciful, I express my sincere appreciation for His boundless guidance, mercy, and blessings that have illuminated my path during this journey, which have been instrumental in enabling me to achieve the completion of this thesis.

My honest thanks to my supervisor, Asst. Prof. Dr. Dhamyaa A. Nasrawi, whose support and expert guidance shaped this thesis. Her encouragement has been invaluable to me.

My heartfelt thanks to Prof.Dr. Noor Dhia Al-Shakarchy for her great support and valuable advice.

I’d like to extend my gratitude to the Computer Science and Information Technology collage's staff, who worked tirelessly to provide us with valuable scientific content.

I owe a debt of gratitude to my family for their patience, support, and endless encouragement, which fuelled my determination.

I also want to thank my dear friends whose support and encouragement contributed to my journey. Their friendship means the world to me, and I'm deeply grateful for their presence in my life.

Haneen Tamim

Abstract

The enormous amount of textual data available around the world, including articles and social media content, has gave the importance for media platforms such as Twitter to employ this data within gender classification. This is an interesting topic for many practical applications such as marketing, recommendation systems and cybercrime. Gender classification in text refers to the process of classifying individuals into one of two genders, male or female, typically based on observable linguistic characteristics.

Natural language processing (NLP) has gained popularity in machine learning field. NLP techniques automatically apply gender classification by using linguistic and stylistic features. This leads to greater participation and satisfaction, improved customer support, and customized content delivery. The dynamic nature and huge vocabulary of the language makes it difficult to determine an author's gender based on linguistic style, it has been a challenge, while the proposed feature extraction method has great importance in overcoming this problem and creating a precise distinction between males and females.

The aim of this thesis is to improve gender classification accuracy based on her/his linguistic style in general domain dataset and specific domain dataset. To achieve this aim, extracting gender textual nuances using text similarity for gender classification improvement and three models have been applied. The first model was applied by three machine learning classifiers Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM) to obtain gender labels and prediction probabilities of them. The second model was applied through the pre-used successful structures of the CNN models to get gender labels and their probabilities. Finally, linear combination mechanism was used by combining additional weights with the label probability results of the previous two models to compute the final prediction probability.

The highest accuracy results of the two datasets (Twitter and TripAdvisor) were acquired. The machine learning model achieved 87.8% on Twitter, while it achieved 75.1% on TripAdvisor. The deep learning model obtained 89.1% on Twitter and 76.3% on TripAdvisor. Finally, the linear combination method achieved (89.6%, 77%) on Twitter and TripAdvisor respectively.

The utilization of the suggested feature extraction technique was crucial in achieving superior outcomes in comparison to earlier researches. In addition, the fact that general domain dataset Twitter's vocabulary is more diverse that helped us outperform TripAdvisor, a specific domain dataset with less accuracy because of its language related only to hotels and restaurants. In order to achieve the maximum accuracy feasible, the using of linear combination strategy utilizing deep learning and machine learning was very important.

Declaration Associated with This Thesis

1. The research paper under the title "Gender Classification Based on Linguistic Analysis: A Review", Haneen Tamim Abd Ali and Dhamyaa A. Nasrawi is presented at the 1st International Conference on Artificial Intelligence Horizons (ICAIH2024) in Sonipat, India, 2024 (Accepted).

2. The research paper entitled: "Extracting Gender Textual Nuances Using Text Similarity for Gender Classification Improvement", Haneen Tamim Abd Ali and Dhamyaa A. Nasrawi at 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA 2024) in Indonesia (Accepted).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Author profiling
BFTree	Best First Tree
BoW	Bag-Of-Words
C-Bi-GRU	Convolutional Bidirectional Gated Recurrent Units
C-Bi-LSTM	Convolutional Bidirectional Long-Short Term Memory
CNN	Convolutional Neural Network
DL	Deep Learning
DNA	Deoxyribonucleic Acid
EDGAD	Egyptian Dialect Gender Annotated Dataset
EEG	Electroencephalography
eWOM	Electronic Word-of-Mouth
FFL	Feminine Frequency Lexicons
GI	Gender Identification
IDH	Inline Digital Holography
KNN	K-Nearest Neighbor
LC	Linear Combination
LBP	Local Binary Patterns
LR	Logistic Regression
LSTM	Long Short-Term Memory

MFL	Masculine Frequency Lexicons
ML	Machine Learning
MLP	Multi-Layer Perceptron
MNB	Multinomial Nave Bayes
MSF	Most Similar Feature
NB	Naive Bayes
NBM	Naïve Bayes Multinomial
NLP	Natural Language Processing
NN	Neural Network
OW	Opposite Weight
PAN	Primary Account Number
POS	Part-of-speech
PPM	Prediction by Partial Matching
ReLU	Rectified Linear Unit
RF	Random Forest
RNN	Recurrent Neural Network
SDTwittC	Saudi Dialect Twitter Corpus
SF	Stylometric Features
SGD	Stochastic Gradient Descent
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine

CHAPTER ONE

INTRODUCTION

1.1 Overview

Generally, the vast majority of the data in the world is textual, encompassing content published on online communities, articles, messages, journals, novels, websites, and other kinds. Text classification, clustering, authorship analysis, sentiment analysis, document summarization, and entity relation modeling are only a few of the tasks involved in NLP [1].

The process of classifying individuals into gender categories, typically male or female, based on various features or attributes it refers to the gender classification [2]. Biological and social data sources, such as text, voice, and image, are used in the complex process of gender classification [3].

Gender classification uses NLP to identify patterns and characteristics in a text, determining an author's gender. As social media usage increases, marketers and governments are increasingly interested in this technique, which can be used for various purposes, such as prediction, crime investigation, security, recommendation [4], marketing, and advertising. In addition, gender classification of authors can help identify cybercrimes by identifying the suspect's gender [5].

Gender classification of authors, as used in machine learning (ML), is the process of automatically classifying people according to similar qualities that are taken from the textual data [6].

Determining the source of textual material on the internet by using simply the data itself is a hurdle. This language challenge is difficult. However, the anonymity and lack of responsibility associated with the internet increase the complexity of the issue by making it simple for anybody to put anything online and falsely claim authorship of it [7].

1.2 Problem Statement

The task of determining the gender of an author based on linguistic style presents a significant challenge due to the multifactorial nature of language. Factors such as cultural heritage, education, personal experiences, and individual mannerisms all contribute to the unique stylistic features of an author's writing. Moreover, the dynamic and evolving nature of language adds further complexity, as authors with a rich and diverse vocabulary introduce nuanced linguistic variations. This requires continuous adaptation to new words, phrases, and expressions, making the task even more intricate. A critical limitation in this area is the insufficient availability of comprehensive and diverse datasets, which hinders the development of accurate and generalizable gender classification models.

1.3 Aims of Thesis

This thesis comprehends the following aims:

1. Gender classification (male/female) based on his/her linguistic style.
2. Improve the classification accuracy in a general domain dataset and a specific domain dataset.

1.4 Objectives of Thesis

To accomplish the above aims, a new gender classification feature extraction method is proposed and has been explained in details:

1. Design a new method of extracting masculine and feminine feature categories to improvement classification accuracy.
2. Present a gender classifier system for a general domain and specific domain using ML algorithms.
3. Present a gender classifier system for a general domain and specific domain using Deep learning convolutional neural networks (CNNs).
4. Present a linear combination (LC) method of combining both ML and Deep learning to improvement classification accuracy.

1.5 Related Works

This section introduces previous studies relevant to the thesis work based on the data input for the classification. There are various forms, such as handwriting [8], names [9], and other forms of media, such as image [10], and audio [11], which are not included because they are considered unrelated.

The main objective of this section is to review the latest studies available on gender classification in textual content (comments, reviews, articles, etc.). It attempts to clarify the datasets, languages, features, methodologies, and metrics used in these studies.

The field of NLP is advancing through machine and Deep Learning (DL) techniques that focus on gender classification within texts. Following studies can be divided into just linguistic analysis (without AI) and gender classification techniques (with AI).

1.5.1 Linguistic Analysis Studies

The study of language and gender is a multidisciplinary field investigating the complex relationship between language and societal attitudes focusing on linguistic patterns and communication styles. The following studies were based on linguistic analysis:

(Patricia R. Owen and Monica Padron, 2016) examined gendered language in action figure narratives for boys and girls, revealing that gender roles were influenced by linguistic elements. Female action figures used social terms, strong adverbs, and trivial references, whereas male narratives reinforced traditional masculinity ideas, including power, aggression, action, and participation in real-world endeavors. These narratives also use second-person plural pronouns, violence terms, and references to disruptive activities, science, and technology [12].

(Mike Thelwall, 2018) examined the influence of gender on sentiment analysis precision using TripAdvisor restaurant and hotel reviews. He used SentiStrength, a lexical sentiment analysis algorithm with machine learning, to analyze gender-specific phrases. He found that females are more likely to display stronger emotions than males, potentially leading to poor marketing decisions [13].

(E .Teso et al., 2018) focused on content created by consumers or electronic word-of-mouth (eWOM) communication on ciao.co.uk, a website for consumer opinions. The research focused on the book genre and used linguistic dimensions, sentiment, and content analysis to identify gender disparities in discourse and preferences between men and women [14].

The research by (Judith C French et al., 2019) found that gender-based differences in the language and writing style of general surgery residency applicant letters of recommendation were insignificant. Both male and female applicants used identical descriptive language and lengths. Linguistic analysis software was used to compare word counts and languages [15].

(Saad Awadh Alanazi, 2019) identified psychometric and stylometric characteristics for gender identification (GI) in the Saudi Dialect Twitter Corpus (SDTwittC) datasets. Word-based qualities are most helpful in resolving GI problems. The results show that Saudi men differ from Saudi women in communication aspects, such as intensifiers, hedges, color, emotion, reason, emoji, and impoliteness. They also differ in politeness, such as greeting, thanking, apologizing, congratulating, encouraging, and best wishing [16].

(L. Balachandra et al., 2021) explored how women used gender-similar language affected investor decisions in venture pitches. It found that female business owners used terminology similar to their male counterparts, with masculine verbal styles being more successful. However, extreme masculine language could be detrimental to both male and female business owners [17].

Table 1.1 presents the overview of the Linguistic analysis studies.

Table 1.1: Overview of Linguistic Analysis Studies

Author	Year	Focus	Method	Findings
Patricia Owen [12]	2016	Gendered language in narratives for action figures.	Analysis of narratives for action figures marketed to boys and girls.	Female narratives contained greater usage of social terms and adjectives related to triviality, while male narratives confirmed traditional masculinity ideas.
Mike Thelwall [13]	2018	Influence of gender on sentiment analysis precision.	Analysis of sentiment in TripAdvisor reviews.	Females displayed stronger emotions than males in reviews, potentially leading to poor marketing decisions.
E. Teso [14]	2018	Gender disparities in consumer-generated content.	Analysis of consumer opinions on ciao.co.uk.	Identified gender disparities in discourse and preferences in book genres, achieving.
Judith French [15]	2019	Gender-based differences in general surgery residency letters.	Linguistic analysis of general surgery residency applicant letters of recommendation.	Found insignificant gender-based differences in language and writing style in applicant letters of recommendation.
Saad Alanazi [16]	2019	Psychometric and stylometric characteristics for GI.	Analysis of the (SDTwittC) dataset.	Identified psychometric and stylometric characteristics for automatic GI, including differences in language use between Saudi men and women.
Balachandra [17]	2021	Impact of gender-similar language on investor decisions.	Study on how gender-similar language affects investor decisions in venture pitches.	Found that female business owners using terminology similar to males had more success.

1.5.2 Gender Classification Techniques Studies

The problem of automatically classifying texts based on the author's gender was addressed by (Aleksandr Sboev et al., 2016). They employed a preexisting corpus of texts in the Slavic language of Russian that had already been RusPersonality-labeled with details on their authors (gender, age, etc.). They aimed to investigate whether it was possible to use ML techniques on relatively context-free criteria to automatically categorize Russian written texts according to their authors' gender. The generated classification models (such as CNN+ Long Short-Term Memory (LSTM)) exhibited at least good accuracy and sometimes even better (up to 0.86 in accuracy, 86% in F1-score) [18].

(Filho et al., 2016) focused on performing the task of gender categorization by extracting gender expression linguistic cues from tweets published in Portuguese utilizing 60 textual meta-attributes, which were frequently utilized on text attribution tasks. The authors used three different machine-learning algorithms (Multinomial Nave Bayes - MNB, Best First Tree - BFTree, and Support Vector Machine - SVM) to classify the author's gender by taking into account characters, structure words, syntax and morphology of short length, multi-genre, content free texts posted on Twitter. It achieved an accuracy rate of 81.66% [19].

(Alsukhni and Alequr, 2016) attempted to use a variety of classification techniques, such as K-nearest neighbors (KNN), J48 decision tree, SVM, MNB, and Nave Bayes (NB), to ascertain the gender of a tweet's author in Arabic. The accuracy of the Naïve Bayes Multinomial Classifier (NBM) was

62.49% without preprocessing and 61.27% with preprocessing. Findings indicated that tweet author names can increase accuracy to above 98% [20].

(Mukherjee and Bala, 2017) suggested a system that uses the maximum entropy and NB algorithms to categorize unstructured text input according to men's and women's preferences for different genders. After analyzing a variety of feature sets, such as voice n-grams, function words, and frequent content terms, they were able to classify gender in microblogs like Twitter with an accuracy rate of 71% [21].

(Alsmearat et al., 2017) examined the GI problem in Arabic articles by applying classifiers from SVM, NB, and Bayesian networks. They employed the bag-of-words (BoW) technique for linguistic analysis and Stylometric Features (SF) to capture writing style characteristics. The results demonstrated that the SF technique, which was less expensive to train, had more accurate results; the top accuracy levels were 80.4% and 73.9%, respectively. The dataset was manually collected from multiple Arabic news websites, such as alrai.com, addustour.com, and sawaleif.com [1].

(Altamimi and J. Teahan, 2017) were interested in solving NLP-related issues with compression-based were interested in using compression-based language models, including Prediction by Partial Matching (PPM), to solve NLP-related problems. The use of PPM for text classification of Arabic text was the focus of the study. When classifying tweets in Arabic text belonging to specifically selected Twitter users, the authors concentrated on authorship and gender classification. They employed gender classification to identify tweets coming from a male or female writer. Their contributions were a comparison of the same data with various ML methods and PPM. They applied a range of machine learning techniques, such as MNB, which produced an accuracy of 92.9%; KNN, which produced an accuracy of 44.4%; SVM, which implemented with an accuracy of 93.9%. When it came to authorship and gender, PPM reached 90% and 96% accuracy, respectively, which was far better than any other ML approach [22].

(Martinc et al., 2017) detailed their strategy for the 2017 Author Profiling (AP) shared work of Primary Account Numbers (PAN), which included developing a model to determine the language and gender preferences of Twitter users. Their suggested logistic regression (LR) classifier's main characteristics were various character and word n-gram types. Additional capabilities include word lists for linguistic variety, emoji and document emotion data, character floods, and part-of-speech (POS) n-grams. Their model achieved accuracy scores of 86% and 98.38%, respectively, in tasks predicting gender and linguistic variation in the Portuguese test set, outperforming all other models. In contrast to test sets for other languages, Spanish (81.93%) and English (80.71%), Arabic (80.31%) had the lowest accuracy [23].

(Bayot and Gonçalves, 2018) evaluated the effectiveness of handcrafted features for predicting age and gender from a collection of texts against the use of CNN and word2vec word embeddings. The network that was developed consisted of a max-over-time pooling layer, dropout layer, embedding layer, SoftMax layer, and convolutional layer. The network was trained to classify the age and gender of the tweets written in Spanish and English. The results showed that age and gender could be classified with the highest accuracy in Spanish (56.0% and 69.3 %, respectively) and English (49.6% and 72.1%, respectively) [24].

The study by (Bsir and Zrigui, 2018) aimed to predict a characteristic label for anonymous text using recurrent neural networks (RNNs) for GI. They used the LSTM neural network architecture and a tweet dataset, achieving an accuracy score of 79.23% [25].

(Park and Woo, 2019) gathered information from Healthboards.com's AIDS discussion forum. They suggested a model for gender detection at took words and emotions out of text messages and used ML, including DL and sentiment analysis, to determine the gender. They classified AIDS patients' gender awareness and developed a learning algorithm that used gender data to find unstructured gender data. They discovered through the experiment that sentiment features produced little accuracy. Sentiment-based terms, however, performed better with the SVM classifier. Classification accuracy was 60.86% SVM, 58.66% Random Forest (RF), and 58.33% NB. Overall, the female group was misclassified frequently by conventional ML methods. With over 90% accuracy, the DL algorithm overcomes this flaw [26].

(Felipe et al., 2020) deal with a cross-domain gender classification job (i.e., AP) based on four domains (Facebook, blogs, e-government requests, and user-generated content) in Brazilian Portuguese. Using word and psycholinguistics-based characteristics, they trained two straightforward classification models—LR and multi-layer perceptron (MLP)—on text. In two different cross-domain situations, their outcomes were compared: first, using a single text source as training data for each task and then merging numerous sources. The F1-scores achieved on Facebook were 80%, opinion was 74%, blog was 78%, and e-government was 79% [27].

(ElSayed and Farouk, 2020) used Neural Network (NN) models to examine the GI (male or female) of followers using the Egyptian dialect. CNN, Artificial Neural Network (ANN), LSTM, Convolutional Bidirectional Long-Short Term Memory (C-Bi-LSTM), and Convolutional Bidirectional Gated Recurrent Units (C-Bi-GRU), which was optimized for the GI problem. The highest-performing model was the C-Bi-GRU model, whose GI accuracies for Egyptian Dialect Gender Annotated Dataset (EDGAD) and PAN'AP 17 might reach 91.37% and 83.7%, respectively [28].

(Vashisth and Meehan, 2020) explored using tweets and NLP methods to classify people by gender. It used vectorizing tweet text, extracting features using BoW, Word Embedding, and conventional ML classifiers. The dataset was publicly accessible for future investigation. Word embedding models outperformed other ML techniques with an accuracy of 57.14% using LR [4].

(Hilte et al., 2022) examined the writing of Flemish teens in private social media communications during interactions between people of the same and different genders. It looked into whether girls and boys write in a more "male" and "female" manner. The two prototype markers of informal online writing that were the subject of the investigation were "oral" and "expressive typographic markers." The findings demonstrated that in interactions with people of different genders, men and women took a more similar tack, with men firmly aligning with a "female" writing style. It achieved 91% average recall and precision [29].

(Koch et al., 2022) found differences in language in WhatsApp messages based on gender and age. Using a sample of 309,229 WhatsApp talks from 226 volunteers, they discovered apparent linguistic disparities related to age and gender. In order to forecast volunteers' age and gender beyond baseline values, they employed cross-validated machine-learning algorithms to determine which language characteristics were the most predictive. The work emphasized methodological approaches for predicting from small and unbalanced text data sets, implications for psycholinguistic theory, and prospective applications of AP. It also drew attention to the mounting threats to individuals' privacy rights. It achieved 85.7% accuracy [30].

(Ikae and Savoy, 2022) examined the efficacy of 10 ML algorithms for identifying gender stylistic distinctions in English web-based communications. It aimed to provide a two-stage feature selection method that reduced feature size to a few hundred terms without significantly affecting performance. Results showed that NN or RF were the best, and the study revealed specific phrases discriminating between genders. It had an accuracy of 82.26% with the MLP classifier [31].

(Onikoyi et al., 2023) concentrating on Twitter, examined the behavioral differences between male and female social media users. They proposed a method for user gender categorization that used NLP and ML techniques to analyze a user's tweets, a profile description, or a short "about me" paragraph. For testing and assessment, they enhanced the Twitter User Gender Classification dataset. To transform text into vectors, they employed a variety of methods and resources, such as the BoW model and pre-trained word embeddings (GLOVE, BERT, GPT2, and Word2Vec). NN, RF, Decision Tree, LR, SVM, XGBoost, Bagging, and Voting Ensemble Classifiers—which used both hard and soft voting to account for all of the aforementioned strategies—were among the popular ML techniques they employed. The best-performing ML algorithm combined with GloVE was RF, achieving an accuracy of 70% [32].

Table 1.2 presents the overview of the Gender Classification Techniques Studies.

Table 1.2: Overview of Gender Classification Techniques Studies

N	Ref	Dataset	Language	Features	Method	Metrics
1	[18]	RusPersonality	Russian	Morphological (POS features), Syntactical parameters.	CNN, LSTM	Both F1-score and accuracy are 86%.
2	[19]	Tweets	Portuguese	Account characters, syntax, words, structure and morphology of short-length, multi-genre, content-free texts.	BFTree, MNB, and SVM.	Accuracy is 81.66% and precision.
3	[20]	Tweets	Arabic	Adding names of Tweet authors as a feature.	Decision Tree, NB, SVM, NB-M, J48 and KNN.	Accuracy is 62.49% and precision.
4	[21]	Tweets	English	Function words and part of speech n-grams.	NB and maximum entropy algorithms	Accuracy is 71% precision, recall and F-measure.
5	[1]	Arabic articles	Arabic	BoW and SF.	BoW approach.	Accuracy is 80.4%, precision, recall, F- measure and Area Under the ROC Curve.
6	[22]	Tweets	Arabic	Character-based compression models (PPM)	MNB, KNN and SVM	Accuracy is 90%, recall and precision.

7	[23]	PAN 2017 set consists of tweets	Portuguese, English, Spanish, Arabic.	The main features are different types of characters and word n-grams. Also, it includes POS n-grams, emoji, and document sentiment information, character flood.	LR Classifier.	Accuracy is 86% in English.
8	[24]	PAN 2016 dataset of Tweets.	English and Spanish	Examine how word2vec word embeddings may be used with CNN	CNN with word2vec word embedding.	Accuracy 72.1% in English.
9	[25]	PAN@CLEF 2017 of tweets	Arabic	Features may refer to the characteristics or attributes of the written text of the LSTM model.	LSTM.	Accuracy 80.06%.
10	[26]	AIDS-related bulletin board at Health boards.	English	BoW, N-grams, and Sentiment Features.	Machine learning, including DL and Sentiment analysis.	Precision, Recall and Accuracy is 90%.
11	[27]	Facebook, Opinion, Blog and E-gov.	Brazilian, Portuguese	Word-based and psycholinguistics-based features.	LR-LIWC LR-TfIdf MLP-skipgram	F1 score 80%.
12	[28]	EDGAD And PAN'AP 17 Dataset.	Arabic	Using DL.	NN, CNN, long–Short Term Memory Convolution Bidirectional Long-Short Term Memory.	Accuracy 91.37%.

13	[4]	Twitter	English	Word Embedding (W2Vec, GloVe), Bag of Words (Term Frequency –Inverse Document Frequency).	Bag of Words, Word Embedding, and LR in this context.	Accuracy 57.14%.
14	[29]	Corpus of private social media messages	Dutch	Expressive typographic markers (e.g., emoticons) which can be considered more ‘female’ features and ‘oral’, speech-like markers.	Best models for oral and expressive Writing.	Precision and recall are 91%.
15	[30]	WhatsApp instant messages	German	General message characteristics and emoji preferences). Also, investigate user demographics.	Cross-validated ML algorithms.	Accuracy 85.7%.
16	[31]	7 CLEF-PAN collections (tweets)	English	Function Words, N grams, Stylistic Markers, POS Features.	10 machine ML.	Accuracy 82.26%.
17	[32]	Twitter	English	Pre-trained word embeddings with the BoW model.	Machine Learning Approach.	Accuracy 70%.

1.6 Thesis Organization

The thesis is divided into five chapters. Each chapter begins with a short background. The summaries of the chapters are as follows:

Chapter Two: This chapter details the theory used in this thesis.

Chapter Three: This chapter focused on the proposed method.

Chapter Four: In this chapter, the results are obtained and discussed.

Chapter Five: This chapter involves conclusions and suggestions for future works.

CHAPTER TWO

THEORETICAL BACKGROUND

2.1 Overview

This chapter presents the details of gender classification techniques and the theoretical background for this thesis, including an overview of gender classification and its approach. In addition, applications for gender classification in author’s text are illustrated.

ML and DL techniques used are demonstrated in this work. Furthermore, some important concepts such as NLP, spaCy library, Counter package, cosine similarity and linear combination are included in this chapter. Finally, the measures used to evaluate the performance of the proposed system are presented.

2.2 Gender Classification

Artificial intelligence (AI) has increasingly become common in numerous facets of daily life. It relies significantly on classification techniques that enable machines to learn from data and make decisions based on that information. In many AI applications such as speech recognition, NLP and image recognition, it is essential to categorize data effectively [33].

Identifying gender is a vital part of the everyday life. Online retail and e-commerce are regions that use gender classification of authors and an increasing number of user engagement and conversion costs [34]. To enhance user pleasure and experience, chatbots and virtual assistants in customer service and guide structures utilize gender classification algorithms to alter their replies and interactions in step with the person's gender [35].

Author’s gender classification in text is a binary or two-class problem that aims to classify anonymous messages or text into one of the specified classes (male or female). Machines find it difficult to assess this task, whereas humans find it simple. People may frequently quickly and easily discern a person's gender class through inspection. Text messages or language phrases are the input of the systems. NLP machine learning usually involves categorizing a series of tokens, such as sentences or documents [4].

The gender classification data can be customized for various purposes, including aiding in business applications such as digital marketing, addressing social scientific inquiries, identifying and prosecuting crimes, defending against terrorism, and preventing embezzlement, falsification [19], advertising, legal research, and understanding societal perceptions of different genders [21].

Gender classification according to textual content is significant in many fields, including the social sciences and healthcare. This makes customized medication treatment and wise decisions possible. Gender classification algorithms compare textual data using ML and NLP, allowing for improved medical results and hospital treatment tailored to a patient's gender [36]. It advances studies on cultural approaches, societal expectations, and disparities in gender in the social sciences, thereby increasing our understanding of gender-associated issues [37].

In recent years, gender classification research has mostly employed visual features instead of textual features [38]. Several state-of-the-art methods can be used to extract meaningful textual characteristics for the gender classification problem, allowing for the advantage of important information found in text materials. Data on gender classification may be in many types such as images, voice and text.

Gender classification in images is a method of using computer algorithms to identify facial features and body characteristics in images. This technology uses ML to differentiate between male and female subjects [39].(A. Khan et. al., 2021) used a method for authentication when evaluated on the dataset, uses SVM with great accuracy and low computing cost for gender classification from iris scans [6].(Thonglim et. al., 2024) used an inline digital holography (IDH) approach to efficiently implement a gender classification model based on fingerprint analysis. It is one of the studies that applied gender classification using fingerprint-based convolutional neural [40].

Voice in gender classification is an area of technology that uses sound quality, tone, and frequency analysis to determine the gender of speakers. It trains computers to distinguish between voices that sound more feminine or masculine [41].(Yavuz T. et al., 2020) using male and female voices through determining gender using sound recordings through ANN, aims to identify gender for forensic informatics and efficient operations by utilizing an equal quantity of male and female voice instances in the dataset [11]. (Alnuaim et. al., 2022) struggles to develop speaker-recognition algorithms because various datasets and features have distinct effects. They also showed how well ResNet50 and deep neural networks perform gender classification [42].

Using word choices, linguistic patterns, and writing styles, gender classification in literature utilizes computer approaches to identify the gender of the author in text. Applications for this technique may be found in literary studies, marketing, and social media analysis [43]. (Ö. ÇELİK and A. F. ASLAN, 2019) applying of ML has made it possible for businesses to anticipate client behavior using predictive models, especially on social media. However, every consumer has different preferences, and gender matters a lot. Through data analysis of 30-70 % of organizations, the goal was to estimate gender in Facebook comments. This can assist businesses in recognizing and focusing on certain client interests in order to provide relevant products or services and increase sales [44]. (M. Arshad et al., 2024) examines AP and its use in marketing, forensics, security, and education, among other fields. It creates an ensemble model called ABMRF by combining a RF with AdaBoostM1. ABMRF consistently outperformed other ML algorithms for age and gender classification in the study using measures such as accuracy, precision, recall, F-measure, and Matthews Correlation Coefficient. The findings demonstrate how well ensemble techniques improve the accuracy of AP tasks, particularly those involving age and GI [45].

2.3 Gender Classification Approaches

There are two types of gender classification studies: appearance- and non-appearance-based. While the non-appearance-based strategy employs biological characteristics and information from the human social network, such as biometrics, bio-signals and blogs, for gender classification, the appearance-based approach determines gender based on external aspects, such as face, steps, and clothing [3]. Figure 2.1 presents gender classification approaches.

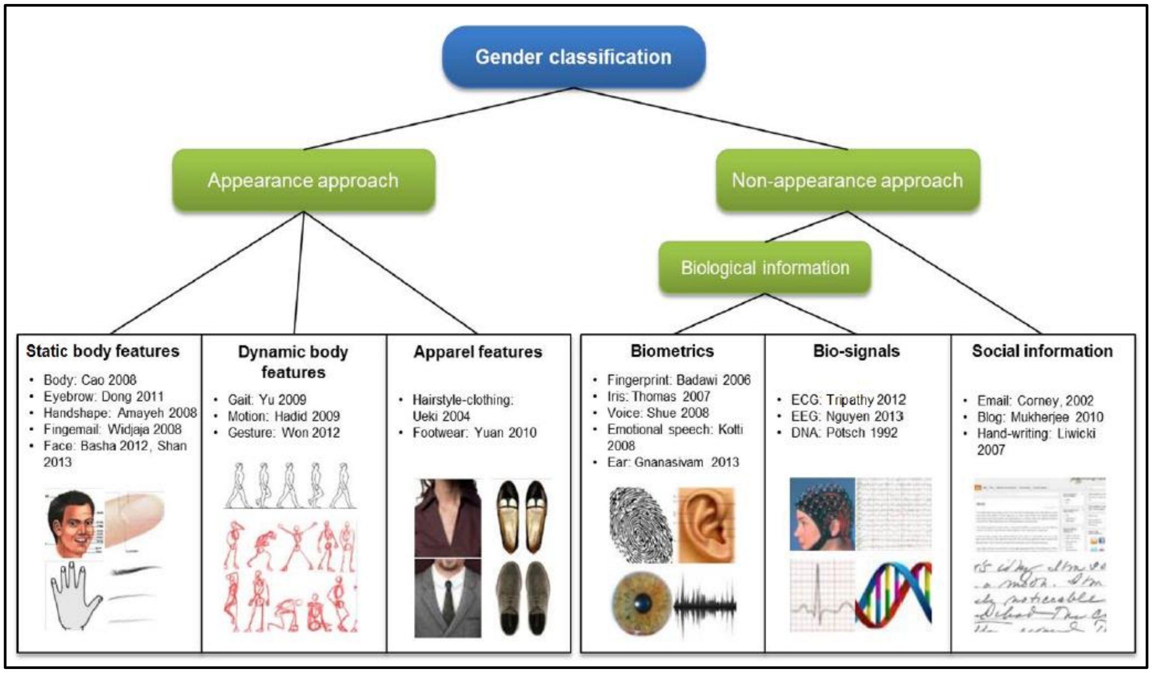


Figure 2.1: Gender Classification Approaches [3]

2.3.1 Appearance Approach

The appearance approach is a strategy based on vision that analyzes a person's physique [46], face [47], eyebrows [48] and fingernails [49], and extracts static, dynamic, and clothing elements [50]. It employs image processing to determine gender; however, because images differ in accuracy, they are not 100% reliable.

2.3.2 Non-Appearance Approach

It extracts features from a person's physical and biometric information, such as voice, iris [6], fingerprint [40] and bio-signals, such as Electroencephalography (EEG) and Deoxyribonucleic Acid (DNA). These features help in gender classification due to differences in language and social style between males and females. In addition to the information collected through a person's daily social interactions, such as email [51], handwriting [52] and blogging [53], is referred to as social network-based information, which it has been used in this thesis.

2.4 Applications of Gender Classification in Author’s Text

Gender classification using of ML techniques performs an essential function in various fields because of its capability to inform decision-making processes [37]. The following are several applications of gender classification in numerous domains.

2.4.1 Security

Worldwide cybersecurity has grown more sensitive to authors' anonymity as a result of the rapid growth of social media platforms (e.g., Twitter, Facebook, and Instagram). By concealing their location, personality, gender, and other personal information online, individuals can avoid being detected by security agencies [16]. Examples in security applications, in which certain companies depend on deception and fraud detection for their survival [1],policy [54],online plagiarism, copyright, and fraud detection investigations [27].

2.4.2 Healthcare

People increasingly use social media to express their experiences, opinions, worries, and beliefs. This leads to the generation of an enormous quantity of useful information that may assist in solving several health-related problems, including health surveillance, mental health, and health care [54].

2.4.3 Forensic

Forensics relies significantly on the automated extraction of information about an author's age, gender, and other demographics from the text. For instance, it would be interesting to understand the linguistic features of an individual who authored a violent text message [18]. Intelligence agencies can use gender classification to investigate cybercrimes [21].

2.4.4 Commercial and Marketing

Classifying people based on their gender allows for a better purchasing experience and advises on marketing efforts. Products may be targeted at certain users via advertising, electronic marketing, and websites. For example, knowing how many male and female consumers are at a supermarket or department store enables business managers to make intelligent choices concerning sales and administration [3].Large companies are interested in the gender and demographic diversity of their product customers based on the analysis of blog entries, especially microblogs. Many business fields, including target advertising and product development, benefit from these evaluations [21]. Companies can use blogs and online product reviews as sources of analysis could find it interesting to find out what kinds of individuals like and dislike their items [18].Other examples include economics and recommendation systems [19].

2.4.5 Literature

Uncertain authorship in works traditionally attributed to male figures, such as Shakespeare, raises critical questions regarding gender classification in literary history. Historically, women and other marginalized groups were frequently compelled to publish anonymously or under pseudonyms, thereby obscuring their contributions and challenging the prevailing male-dominated narratives of authorship. The intricate and nuanced portrayal of female characters within Shakespeare’s oeuvre, for instance, suggests the possibility of collaborative or unacknowledged contributions from women. This complicates the assumption of singular male authorship and invites a reevaluation of the gendered dynamics that have shaped literary production and recognition [1].

2.5 Natural Language Processing in Gender Classification

Recently, there has been an increase in interest in the use of NLP for gender classification of textual information. The use of NLP techniques is beneficial in a variety of fields, including gender studies, computational linguistics, and sociolinguistics. These approaches enable automatic structures to evaluate and categorize textual materials based largely on gender development [55].

One established method for determining gender is to use stylistic and linguistic signals found in the text. An investigation demonstrated that variances in language usage between both genders, together with preferred terminology, sentence structure, and conversation patterns, can serve as accurate markers for gender classification [56].

Nonetheless, there are challenges and moral issues surrounding gender classification in textual material. Stereotypes and gender disparities established in training data can lead to biased predictions and reinforce existing disparities in society using NLP techniques [57].

Combining NLP with gender classification yields immediate benefits in a wide range of domains. The investigation indicated that it could effectively identify gender bias in online conversations, enhance advertising and marketing strategies, and enhance human reviews. NLP speeds up the selection process by automating gender classification in textual materials. This will enable well-timed interventions to reduce gender bias and enhance communication efficacy. This relationship enables customized content material delivery and improves customer service interactions, which raises levels of satisfaction and engagement [58].

2.6 NLP Types with Gender Classification in Author’s Text

Textual gender classification improves accuracy and robustness through the use of various NLP techniques to determine gender from text. Some types on NLP are discuses in the following.

2.6.1 Lexical-based Classification

Classification based on gender relies on the ability to arise the gender of individuals described in a text by analyzing how language works such as vocabulary, grammar, and conversation styles. This approach uses words and phrases that are often associated with specific genders to classify textual data. For instance, (Kozlowski et al., 2019) investigated methods based only on gender class and lexical analysis in social media data [59].

2.6.2 Stylistic-based Classification

This model examines writing styles and patterns that are specific to a person’s gender. It considers the structure of the sentence row and the task of speech. Research by (Burger et al., 2019) compared stylistic features in the text of males and females and evaluated the effectiveness of predicting gender [60].

2.6.3 Machine Learning-based Classification

Utilize acquired algorithms and statistical models that have been learned with labeled training data to forecast gender characteristics in the text. Their models capture regularities and links between types of textual evidence and gender tags, enabling quick and accurate gender categorization of new information. Therefore, (Bergsma et al. ,2013) they explored ML methods for Twitter data gender categorization and achieved strong output with feature-based and DL systems [61].

2.6.4 Hybrid Approach

Several hybrid approaches combine lexical analysis, stylistic indications, ML algorithms, and other techniques to achieve improved robustness and accuracy in textual GI. Combining their strengths and weaknesses enables the creation of a gender prediction model that is more reliable. In an online forum, (Nguyen et al., 2016) utilized a hybrid gender categorization framework to show how different approaches might be merged to provide superior outcomes [62]. This type, which combines many techniques, is the one that this thesis employs.

2.7 Machine Learning

The scientific study of algorithms and statistical models that computer systems employ to perform a particular task without being explicitly programmed is known as machine learning. Learning algorithms for a variety of everyday applications. One of the reasons a learning algorithm that has learned the art of ranking websites makes a search engine, such as Google, perform so effectively every time it is used to search the Internet. These algorithms are employed in many fields, including predictive analytics, image processing, and data mining. The primary benefit of ML is that algorithms can operate autonomously once they determine what to do with data [63].

At this point, classification, analysis, and recommendation are the different subdomains of ML. It has been effective in the implementation of document classification, image analysis, medical diagnosis, network intrusion detection prediction, and denial-of-service attack prediction [64]. One type of ML task that is frequently used in gender prediction tasks is supervised learning algorithms, which use labeled data to learn and then classify or predict. A training dataset with each data instance linked to its matching gender label was necessary for these algorithms [65].

2.7.1 Types of Machine Learning Techniques

Various machine learning techniques can be applied to analyze data in a specific problem domain and extract insights or useful knowledge for creating intelligent applications for the real world. Machine learning techniques can be divided into three major categories based on the provided indicators used to analyze and understand the situation, as illustrated in the following. Figure 2.2 presents the types of machine learning techniques.

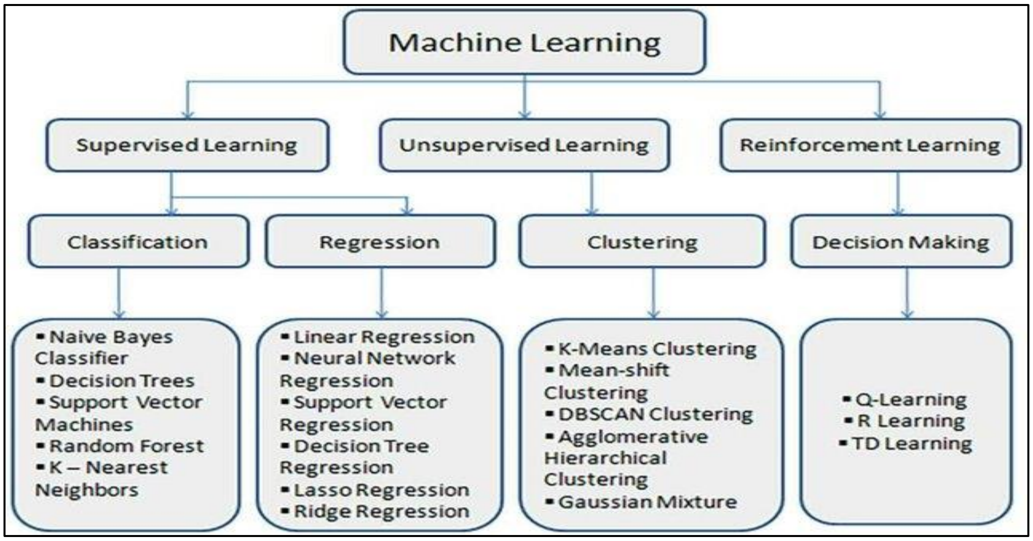


Figure 2.2: Machine Learning Techniques [66]

1. Supervised Learning

Data that has been labeled and divided into two sets—a testing data set and a training data set—is required for supervised machine learning algorithms. There are certain outputs from the learned data set that require prediction. The goal is to train the computer to recognize patterns from the training data set that are similar to those found in the test data set and use those patterns to predict the real-valued output [67].

The two most popular supervised tasks are data fitting (called "regression") and data separation (called "classification"). One example of supervised learning is text classification, which is the process of predicting the class label or sentiment of a text segment, such as a tweet or a product review [68]. In this thesis, it has been used supervised learning techniques and classification task.

2. Unsupervised Learning

Machine learning may be done without labeled data by using an approach called unsupervised learning. Based on the input data, an algorithm is created and then tested on a set of data. While the testing data set aids in making accurate value predictions, the training data set is utilized to create and train the model. The algorithm learns from the previously supplied information and forecasts the real-valued outcome by drawing on prior experiences [67].

Unsupervised learning is practically the same as clustering. Due to the lack of class labeling in the input instances, the learning process is unsupervised. Clustering is typically used to find classes in the data. An unsupervised learning technique, for instance, can be fed a collection of photos showing handwritten numbers. Assume that ten data clusters are discovered. These clusters may represent the ten different numbers, 0 through 9, in that order. Nevertheless, the learnt model is unable to provide the semantic significance of the clusters discovered since the training data are not labeled [69].

3. Reinforcement Learning

Since the algorithm merely receives a response indicating whether the output is accurate or not, reinforcement learning is thought of as an intermediate kind of learning. To reach the right result, the algorithm must investigate and rule out a number of options. As the algorithm makes no recommendations or fixes for the issue, it is referred to as learning with a critic [70].

The goal of reinforcement learning science is to help living things make the best choices and exhibit reward-motivated behavior. Understanding the essential elements of a learning agent's interaction with its surroundings in order to accomplish a target is its goal. An environment-related purpose and the ability to detect the state of the environment are requirements for the agent. Machine learning models are trained through reinforcement learning to make judgments in complicated situations. When faced with a situation similar to a game, the agent experiments to discover a solution. Enhancing the machine's overall intelligence is the aim of increasing the reward for its activities [71].

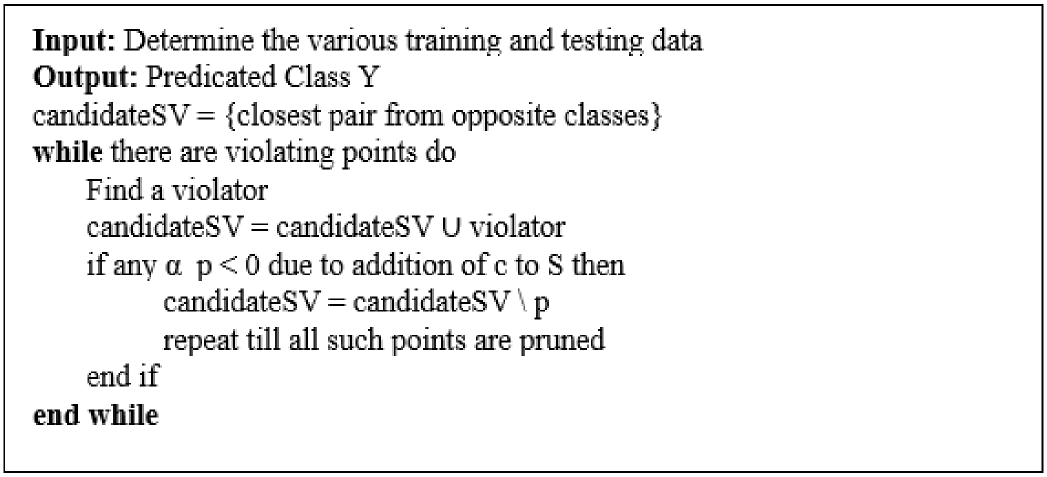
2.7.2 Machine Learning Algorithms

The principles of three ML algorithms (SVM, RF and LR) that are used are discussed in the following.

1. Support Vector Machine

In pattern classification, the development of a model that maximizes the performance given the training data is the main goal. Using traditional training techniques, the models are chosen to ensure that every input-output combination is appropriately classified into the appropriate class. However, the model starts to memorize the training data instead of learning to generalize, which weakens the classifier's capacity to generalize if the classifier is well matched to the training set. SVM's main objective is to divide the classes of a training set using a surface that optimizes the distance between each class. In other words, SVM makes it possible to maximize a model's capacity for generalization [72].

SVM is currently utilized for a wide range of tasks, including feature selection, regression estimation, multiclass classification, and binary class pattern recognition [73]. Figure 2.3 present pseudocode of SVM.

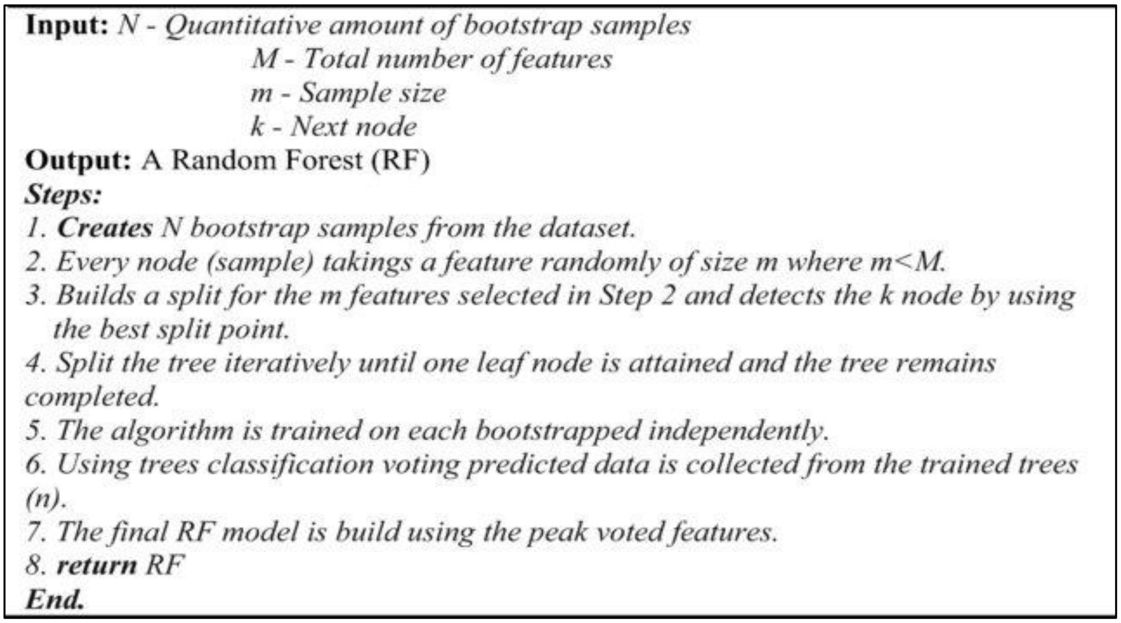


```
Input: Determine the various training and testing data
Output: Predicated Class Y
candidateSV = {closest pair from opposite classes}
while there are violating points do
    Find a violator
    candidateSV = candidateSV ∪ violator
    if any $\alpha$  p < 0 due to addition of c to S then
        candidateSV = candidateSV \ p
        repeat till all such points are pruned
    end if
end while
```

Figure 2.3: Pseudocode of Support Vector Machine [74]

2. Random Forest

An ensemble-based learning system called random forest is composed of many collections of correlated decision trees. It is based on the notion of bootstrap aggregation, a technique for resampling and replacing data to reduce variance. When predicting, RF employs the multiple regression coefficient (regression) or the classification of majority votes in the final two nodes. Constructed on the concept of decision trees, RF models have produced notable gains in prediction accuracy when compared to singlet trees because they increase the number of trees by n; every training set tree is randomly picked without being replaced .Decision trees are simply composed of a structure similar to a tree, with the top node serving as the tree's root and being split recursively into a number of decision nodes that branch out from there until they reach the terminal or decision node [75]. Figure 2.4 presents pseudocode of RF.



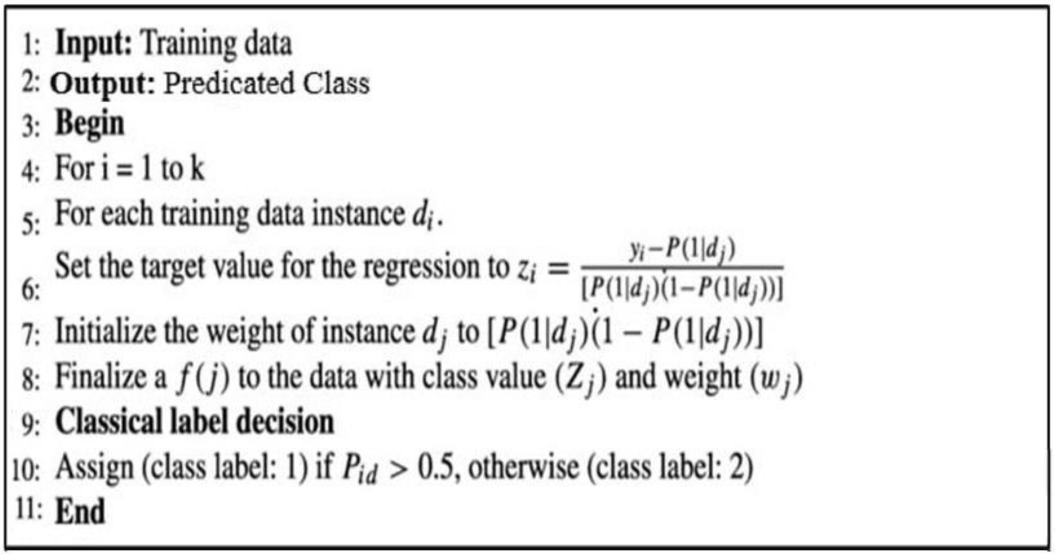
```
Input: N - Quantitative amount of bootstrap samples
             M - Total number of features
             m - Sample size
             k - Next node
Output: A Random Forest (RF)
Steps:
1. Creates N bootstrap samples from the dataset.
2. Every node (sample) takings a feature randomly of size m where m<M.
3. Builds a split for the m features selected in Step 2 and detects the k node by using
   the best split point.
4. Split the tree iteratively until one leaf node is attained and the tree remains
completed.
5. The algorithm is trained on each bootstrapped independently.
6. Using trees classification voting predicted data is collected from the trained trees
(n).
7. The final RF model is build using the peak voted features.
8. return RF
End.
```

Figure 2.4: Pseudocode of Random Forest [76]

3. Logistic Regression

A number of independent variables that describe a link to a dependent response variable make up linear models. Supervised learning in ML terminology refers to mapping qualitative or quantitative input features to a target variable that is intended to be predicted, such as financial, biological, or sociological data, provided that the labels are known. LR is one of the most often used linear statistical models for discriminant analysis [75].

LR may be used to solve classification problems. It returns the binomial outcome, or the chance (between 0 and 1) that an event will occur based on the values of the input variables. Examples of binomial outcomes of LR include determining whether a tumor is benign or malignant or whether an email is regarded as a spam. Multinomial findings, such as a prediction of the favored cuisine—Arabic or Italian, can also be obtained through the use of LR [77]. Figure 2.5 presents pseudocode of LR.



```
Input: Training data
Output: Predicated Class
Begin
For i = 1 to k
For each training data instance $d_i$.
Set the target value for the regression to $z_i = \frac{y_i - P(1|d_j)}{[P(1|d_j)(1-P(1|d_j))]}$
Initialize the weight of instance $d_j$ to $[P(1|d_j)(1 - P(1|d_j))]$
Finalize a $f(j)$ to the data with class value $(Z_j)$ and weight $(w_j)$
Classical label decision
Assign (class label: 1) if $P_{id} > 0.5$, otherwise (class label: 2)
End
```

Figure 2.5: Pseudocode of Logistic Regression [78]

2.8 Deep learning

Recently, DL has gained popularity in the computer industry. It is a subset of machine learning, and is used in many real-time applications. It requires a large amount of data to make decisions regarding new data. Deep Neural Networks (DNN) are a type of NN that is used for data processing. The phrase "deep neural networks" has gained popularity because NN are frequently employed in DL techniques. The CNN is one of the most frequently utilized deep neural networks and does not require human feature extraction, in contrast to traditional feature extraction methods such as Scale-Invariant Feature Transform (SIFT) and Local Binary Patterns (LBP) [79].

In the training process, backpropagation computes the gradient of the loss function to update the model's weights using a gradient-based approach like Stochastic Gradient Descent (SGD) [80],[81]. SGD randomly selects small batches of data for each iteration, which helps escape local minima and speeds up training. If the loss function is convex, SGD guarantees that it will find the global minimum [82]. However, depending on the learning rate and step size, the optimization path may vary. To improve performance and reduce training time, learning rate decay is often used, where the learning rate is initially high to allow large weight adjustments and progressively reduced over time to fine-tune the model. For example, starting with a learning rate of 0.1, it might be reduced to 0.01 after a few epochs, making large adjustments early on and smaller, precise updates as the model approaches the optimal solution [83].

2.8.1 Convolutional Neural Network Architecture

The CNN are widely used algorithms in DL to automatically identify relevant features without human supervision. CNNs are inspired by neurons in human and animal brains, similar to conventional neural networks. They offer three key benefits: sparse interactions, parameter sharing, and equivalent representation. Unlike conventional fully connected networks, CNNs use shared weights and local connections, which simplify the training process and speed up the network [80]. CNNs outperform advanced algorithms in several cases when they are used with text and image datasets [26].

It has been shown to be efficient for NLP applications, including sentiment analysis, language modeling, and text classification. The ability of CNNs to identify local structures in an input is one of their main advantages. This can be related to NLP by collecting n-grams or word sequences that are close to one another, which might be crucial for understanding the meaning of a phrase or document. The ability of CNNs to interchange parameters over a wide range of input regions is an important property. This can help the model generalize new data and reduce the number of parameters required to train the model [80].

The pre-training of CNNs on large amounts of unlabeled data can help improve their performance in subsequent NLP tasks. While conventional NLP techniques such as BoW and n-gram models have proven successful for a variety of applications, CNNs can provide a strong and adaptable substitute that can effectively capture local structures, share parameters, and generalize to new data [84] .

Convolutional neural networks, can recognize and simulate intricate patterns in data by introducing non-linearity through the use of activation functions. Two popular activation functions are Sigmoid or Tanh, which are good for binary classification but may suffer from vanishing gradients in deep networks; Rectified Linear Unit (ReLU) accelerates convergence and solves vanishing gradient issues. Tasks like text categorization and object identification are made more efficient by these functions [85]. An overview of CNN architecture is shown in (figure 2.6).

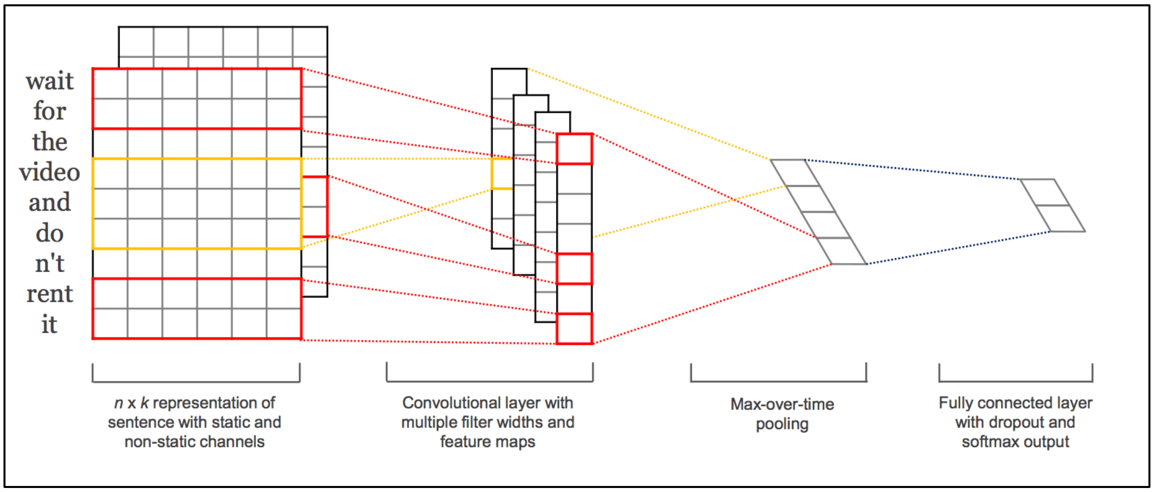

wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 2.6: Overview of Convolutional Neural Network Architecture [24], [26]

The four layers of CNN architecture are demonstrated below [86]:

1. **Input layer:** The entries in this layer are "n". Since each element is represented by a dense vector with size k, the input may be thought of as a feature map with dimensionality k × n.
2. **Convolutional layer:** This layer, which comprises most of the computation process, is most crucial and essential layer of a CNN. The neurons in the convolutional layer are managed by extracting a collection of connected feature maps. This layer is composed of a collection of learnable filters, often known as kernels, which yield two-dimensional activation maps and output volume when stacked and concatenated along the depth dimension. It serves as a symbol for learning from sliding w-grams.
3. **Max-pooling:** The convolutional and pooling layers of the CNN model extract higher-level features and reduce the spatial dimensions without affecting the information. The model's parameter count was optimized to manage overfitting and reduce computational complexity. Pooling processes include multiscale order less, spatial pyramid, spectral, average, stochastic, and max-pooling. The max-pooling layer smoothens and compresses the data, making it invariant to slight translational changes. In max pooling, the maximum value contained in the window is the output, and a predetermined filter is applied across the nonoverlapping subregions of the input. Max pooling reduces dimensionality and the computational cost of learning several parameters [83].
4. **Fully connected layer:** The final layer of the CNN, which is a systematic NN, links every neuron in the forward and preceding layers. A convolution layer cannot follow a completely linked layer because of non-ordered

neurons. Recent designs, like "Network in Network," have replaced the whole connecting layer with a global average-pooling layer. In addition, there is the dropout strategy that works in fully connected layer, which can be applied to deep neural networks to reduce the problem of overfitting. During training, this strategy is implemented by randomly removing units and their connections [87].

2.9 SpaCy Library

An accurate, free, open-source library for advanced NLP parsing using Python. It processes the text for the purpose to comprehends and delineates it, regardless of its size. Moreover, because of its many built-in features, it's a useful tool for language modeling and text processing [88].It offers an extensive feature set for text processing encompassing tasks such as tokenization, POS tagging, and named entity recognition [89].

2.10 Counter Package

Subclass of 'dicts' in Python. Counter[1] class was specifically designed to rely on hashable objects. It provides an easy and sustainable way to generate a frequency distribution or count instances of an item in an iterative manner. Dictionary values are saved for the counts of the elements in the collection, while the elements individually are stored as dictionary keys. This is significant in real situations such as text processing, which counter has used to count the number of times a term appears in a document.

[1] https://docs.python.org/3/library/collections.html#counter-objects

2.11 Cosine Similarity

A measure of the similarity between two vectors in an inner product space is called the cosine similarity. If two vectors are generally pointing in the same direction, they may be determined by measuring the cosine of the angle between them. It is frequently employed in text analysis to measure document similarity. Thousands of features can be used to characterize a document, each of which records the frequency of a certain word or phrase (e.g., a keyword) in the text. Consequently, a term-frequency vector is used to represent each document as an object. These structures are used in biological taxonomy, gene feature mapping, text document clustering, and information retrieval [90]. Eq. (2.1) shows how to calculate similarity.

$$sim(x, y) = \frac{x.y}{||x||||y||} \tag{2.1}$$

Where ||x||, ||y|| are the Euclidean norm of vector x = (x_1, x_2..., x_p), y= (y_1, y_2..., y_p) respectively. A cosine value of 0 means that the two vectors have no match, while the smaller angle means the greater match between vectors [90].

2.12 Linear Combination

A linear combination is a foundational concept in mathematics, particularly in linear algebra, and it is widely used in various fields such as machine learning, optimization, and system theory. The concept refers to constructing new elements (such as vectors or functions) by combining existing ones through scalar multiplication and addition. This method is essential in representing complex systems and solving real-world problems where data or parameters are linearly related [91].Given a set of vectors v_1, v_2, ..., vn in a vector space V, a linear combination of these vectors is expressed in Eq. (2.2):

$$w = c_1 v_1 + c_2 v_2 + \cdots + cnvn \tag{2.2}$$

where:
- v_1, v_2, ..., vn are vectors from the vector space V,
- c_1, c_2, ..., cn are scalars (real or complex numbers),
- w is the resulting vector from the linear combination.

2.12.1 Properties of Linear Combinations

Linear combinations exhibit several key properties that are fundamental in linear algebra and vector spaces. These properties help in understanding how vectors relate to each other and how they span a given space. The most important properties of linear combinations include [92]:

1. Scalars: The constants c_1, c_2, ..., cn are known as the scalars or weights. They determine the contribution of each vector vi in forming the final vector w. When all scalars are zero, the linear combination results in the zero vector.

2. Span of a Vector Space: The set of all possible linear combinations of a set of vectors $\{v_1, v_2, ..., vn\}$ forms a subspace of the vector space. The span of the vectors v_1, v_2, ..., vn is the collection of all vectors that can be expressed as their linear combination: Span $(v_1, v_2, ..., vn) = \{w \mid w = c_1v_1 + c_2v_2 + \cdots + cnvn$, for some $c_1, c_2, ..., cn \in \mathbb{R}\}$

3. Linear Independence: A set of vectors v_1, v_2, ..., vn is said to be linearly independent if no vector in the set can be written as a linear combination of the others. Mathematically, the set is linearly independent if the only solution to: $(c_1v_1 + c_2v_2 + \cdots + cnvn = 0)$ is $(c_1 = c_2 = \cdots = cn = 0)$. If such a solution exists where the scalars are not all zero, the vectors are linearly dependent.

2.13 Evaluation Measures

A model's performance is determined through the use of evaluation measures. There are several metrics that may be used for evaluating a model. Performance metrics come in extremely useful for comparing and evaluating various ML or classification models [93].

2.13.1 Performance Measure Analysis

It is common practice in classification problems to include the expenses of making correct or incorrect classifications. When the cost of many misclassifications varies significantly, this may be beneficial. A (statistical) metric for classification accuracy indicates the effectiveness of the classifier in accurately recognizing items [94] . The accuracy and error rate are below in Eqs. (2.3) and (2.4) [95].

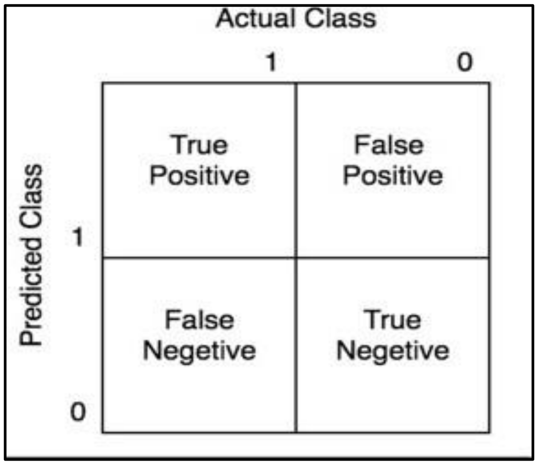
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.3)$$

$$Error\ rate = 1 - Accuray \qquad (2.4)$$

2.13.2 Confusion Matrix

The confusion matrix is a cross-tabulation which records the frequency of occurrences between two raters, as well as the expected and true/actual classifications. Since the classes are arranged in the rows and columns in the same order, the items that have been correctly classified and the number of times the two raters agree are positioned on the main diagonal, which runs from top left to bottom right [93].

The confusion matrix is a scalar statistic used for model comparison or parameter adjustment, allowing for customization for application-specific needs. It can account for class imbalance in data and asymmetric cost weighting of prediction outcomes. The confusion matrix summarizes the prediction performance of a binary classifier, representing the counts of true positives, false positives, false negatives, and true negatives, labeling one class as positive and the other as negative [96].



		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negetive	True Negetive

Figure 2.7: The Confusion Matrix for Binary-Class [97]

2.13.3 K-fold Cross Validation

An ANN model must have an optimal set of parameters, such as the number of hidden nodes and back-propagation learning rate, to be evaluated accordingly. A DL model's dependability estimate may be substandard if the training and testing datasets are unfairly divided such that none of them becomes representative of all data. The k-fold cross-validation process computes the average score to acquire the assessment results for several neural nets. Subsequently, it delivered a performance measure for the entire run. This procedure involved repeating the training and testing K times across different data samples from the collection. This is useful when there are few data samples, but it is computationally expensive. When using this cross-validation method, K different models were applied each time, and one portion of the dataset was retained for evaluation, with the remaining portion being used for instruction. Finally, the findings were combined to obtain the error rate [98]. Figure 2.8 presents cross validation with 5-fold.

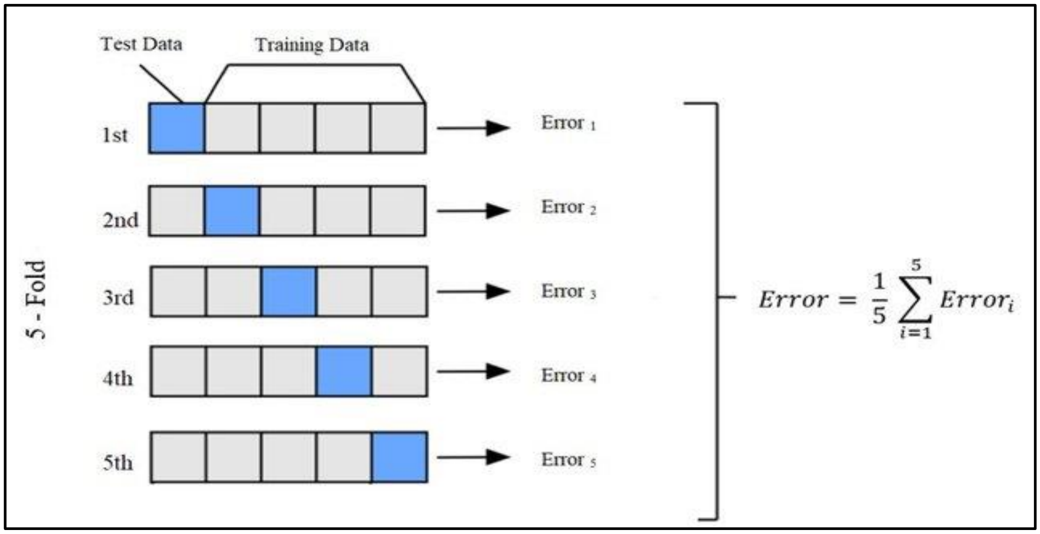


Figure 2.8: Cross Validation with 5-Fold [99]

2.13.4 Cross Entropy

A machine learning model that performs better is achieved by fine-tuning its weights using the cost function during the NN training. In particular, during forward prediction, the NN is trained using training set data, and results are produced that, when it comes to classification, indicate the possibility or consistency of the potential labels. The loss function computes the cost for each variation between the target label and NN outputs after comparing these probabilities to the target labels. For every trainable weight in the NN, the partial derivative of the loss function is computed during the backpropagation. These partial derivatives were used to modify the weights. In typical circumstances, backpropagation iteratively modifies the NN's trainable weights to create a model with less loss [100].In neural network training, the cost function is essential for adjusting weights to produce a better machine learning model. While back-propagation computes partial derivatives for every trainable weight, forward propagation produces probabilities in relation to goal labels. These derivatives modify the weights, and backpropagation iteratively modifies the trainable weights to generate a model with less loss under typical circumstances. A machine learning model that is more precise and effective is guaranteed by this procedure. The binary cross-entropy loss function used for binary classification problems in machine learning is represent in Eq. (2.5) [100].

$$J_{bce} = -\frac{1}{M}\sum_{m=1}^{M}\left[y_m \times \log\left(h_\theta\left(x_m\right)\right) + \left(1 - y_m\right) \times \log\left(1 - h_\theta\left(x_m\right)\right)\right] \quad (2.5)$$

Where M: number of training examples,y_m: target label for training example m, x_m : input for training example m and h_θ: model with neural network weights θ.

CHAPTER THREE

PROPOSED METHOD

3.1 Overview

This chapter details the proposed work, which includes a model that improves the accuracy of the Twitter dataset and determines the accuracy of the TripAdvisor dataset using ML and DL individually. The two learning models are then combined for each dataset, and the accuracy is improved by using linear combination method. Before describing the proposed methodology, referring to the datasets that have been used in the thesis.

3.2 Dataset

Two English datasets are used in this thesis:

- Twitter [4], [32]: the first is a general domain dataset represented by brief phrases include the text of tweets (often capped at 280 characters). It called general domain due to the various topics that have been contained such as educational, artistic, political and health. Twitter contains 20,051 accounts, that comprises of data, such as name, description, gender, tweets and etc. It comprises two types of text (description and tweets) which are used in the thesis. The Twitter dataset is available in CSV format and is ready to work on. It can be found publicly on Kaggle[2].
- TripAdvisor [13]: the second is a specific domain dataset provided detailed reviews of restaurants and hotels in the United Kingdom using long phrases. Its particular theme, which contains information about hotels and restaurants, is why it is considered a specialized domain. It

[2] https://www.kaggle.com/datasets/crowdflower/twitter-user-gender-classification?resource=download

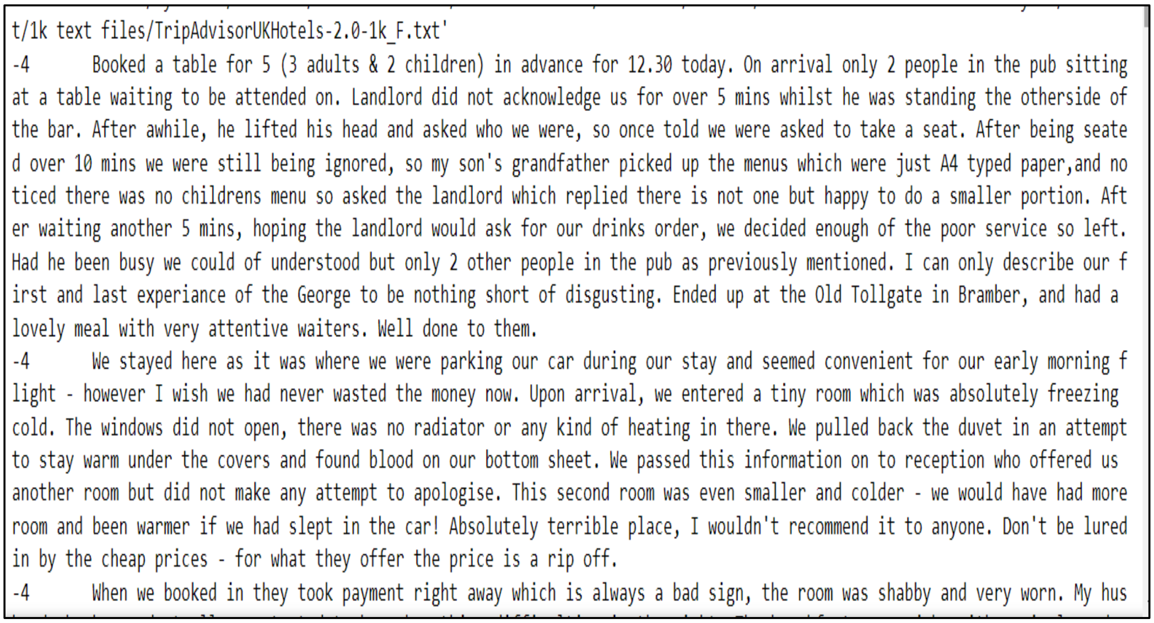
consists of 189,538 reviews, which it is text only with the gender as a label of the text of the dataset. To the best of our knowledge, this is the first time using this dataset to gender-classification domain that focuses exclusively on opinions and textual evaluations of eating and lodging experiences. Prior to commencing work on the TripAdvisor [3], certain steps are performed. Initially, the data is available as text files containing lengthy reviews of restaurants and hotels. Each file has a label indicating either the female or male. These files are converted into data frames, labels (female or male) are added for each row, and then are transformed into CSV files. This constitute the final dataset in which the proposed methodology is applied. Figures 3.1, 3.2 present samples of TripAdvisor dataset before and after the steps of conversion to CSV file. Due to RAM limitations, TripAdvisor dataset has been randomly sampled from its original large-scale version.

Dataset information is shown in Table 3.1, which includes number of records of the dataset, number of records in (females/males), and the average tokens in the dataset. It can be shown that in Twitter there is only 12894 of (females/males), because other labels in gender are organizations or unknown.

Table 3.1: Dataset Information

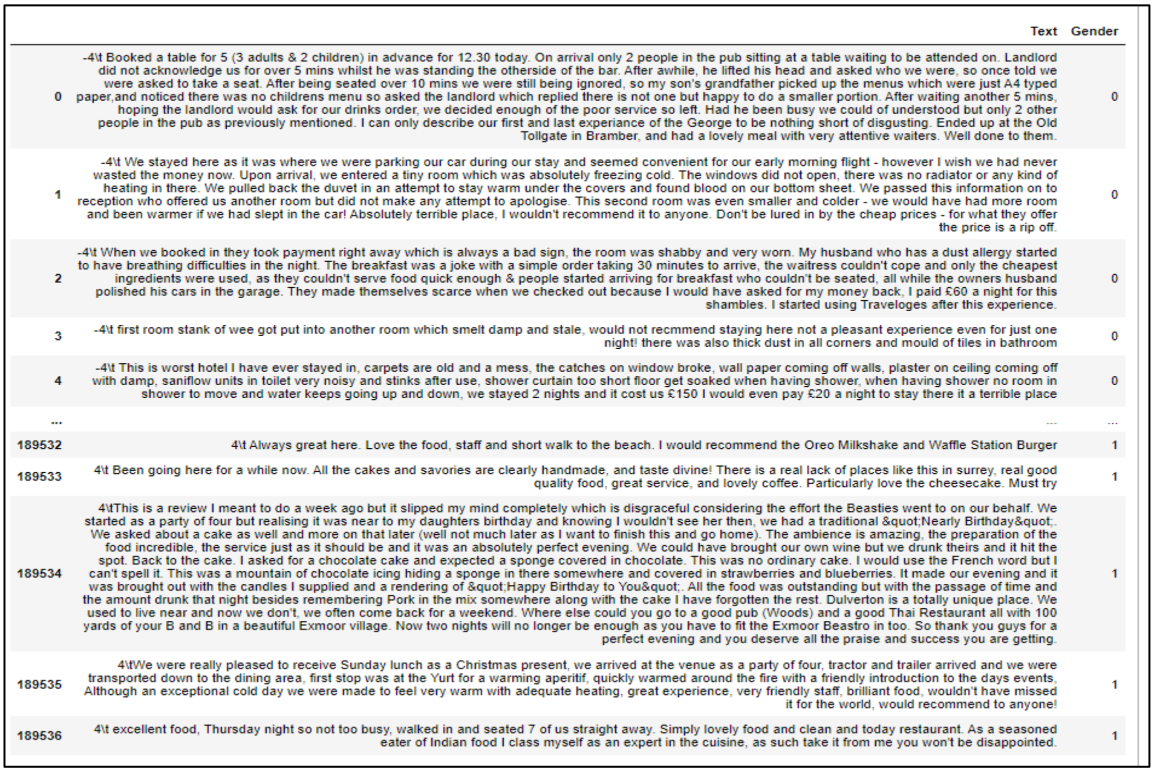
Dataset	**# of Records**	**# of Females**	**# of Males**	**Female %**	**Male %**	**Avg. of Tokens**
Twitter	20050	6700	6194	33.4	30.8	18.37
TripAdvisor	189537	94769	94768	50.01	49.99	148.26
Sampled-TripAdvisor	20000	10000	10000	50	50	147.52

[3] https://figshare.com/articles/dataset/TripAdvisor_reviews_of_hotels_and_restaurants_by_gender/6255284



```
t/1k text files/TripAdvisorUKHotels-2.0-1k_F.txt'
-4	Booked a table for 5 (3 adults & 2 children) in advance for 12.30 today. On arrival only 2 people in the pub sitting
at a table waiting to be attended on. Landlord did not acknowledge us for over 5 mins whilst he was standing the otherside of
the bar. After awhile, he lifted his head and asked who we were, so once told we were asked to take a seat. After being seate
d over 10 mins we were still being ignored, so my son's grandfather picked up the menus which were just A4 typed paper,and no
ticed there was no childrens menu so asked the landlord which replied there is not one but happy to do a smaller portion. Aft
er waiting another 5 mins, hoping the landlord would ask for our drinks order, we decided enough of the poor service so left.
Had he been busy we could of understood but only 2 other people in the pub as previously mentioned. I can only describe our f
irst and last experiance of the George to be nothing short of disgusting. Ended up at the Old Tollgate in Bramber, and had a
lovely meal with very attentive waiters. Well done to them.
-4	We stayed here as it was where we were parking our car during our stay and seemed convenient for our early morning f
light - however I wish we had never wasted the money now. Upon arrival, we entered a tiny room which was absolutely freezing
cold. The windows did not open, there was no radiator or any kind of heating in there. We pulled back the duvet in an attempt
to stay warm under the covers and found blood on our bottom sheet. We passed this information on to reception who offered us
another room but did not make any attempt to apologise. This second room was even smaller and colder - we would have had more
room and been warmer if we had slept in the car! Absolutely terrible place, I wouldn't recommend it to anyone. Don't be lured
in by the cheap prices - for what they offer the price is a rip off.
-4	When we booked in they took payment right away which is always a bad sign, the room was shabby and very worn. My hus
```

Figure 3.1:   Sample of TripAdvisor Dataset Before Conversion to CSV



	Text	Gender
0	-4\t Booked a table for 5 (3 adults & 2 children) in advance for 12.30 today. On arrival only 2 people in the pub sitting at a table waiting to be attended on. Landlord did not acknowledge us for over 5 mins whilst he was standing the otherside of the bar. After awhile, he lifted his head and asked who we were, so once told we were asked to take a seat. After being seated over 10 mins we were still being ignored, so my son's grandfather picked up the menus which were just A4 typed paper,and noticed there was no childrens menu so asked the landlord which replied there is not one but happy to do a smaller portion. After waiting another 5 mins, hoping the landlord would ask for our drinks order, we decided enough of the poor service so left. Had he been busy we could of understood but only 2 other people in the pub as previously mentioned. I can only describe our first and last experiance of the George to be nothing short of disgusting. Ended up at the Old Tollgate in Bramber, and had a lovely meal with very attentive waiters. Well done to them.	0
1	-4\t We stayed here as it was where we were parking our car during our stay and seemed convenient for our early morning flight - however I wish we had never wasted the money now. Upon arrival, we entered a tiny room which was absolutely freezing cold. The windows did not open, there was no radiator or any kind of heating in there. We pulled back the duvet in an attempt to stay warm under the covers and found blood on our bottom sheet. We passed this information on to reception who offered us another room but did not make any attempt to apologise. This second room was even smaller and colder - we would have had more room and been warmer if we had slept in the car! Absolutely terrible place, I wouldn't recommend it to anyone. Don't be lured in by the cheap prices - for what they offer the price is a rip off.	0
2	-4\t When we booked in they took payment right away which is always a bad sign, the room was shabby and very worn. My husband who has a dust allergy started to have breathing difficulties in the night. The breakfast was a joke with a simple order taking 30 minutes to arrive, the waitress couldn't cope and only the cheapest ingredients were used, as they couldn't serve food quick enough & people started arriving for breakfast who couldn't be seated, all while the owners husband polished his cars in the garage. They made themselves scarce when we checked out because I would have asked for my money back, I paid £60 a night for this shambles. I started using Traveloges after this experience.	0
3	-4\t first room stank of wee got put into another room which smelt damp and stale, would not recmmend staying here not a pleasant experience even for just one night! there was also thick dust in all corners and mould of tiles in bathroom	0
4	-4\t This is worst hotel I have ever stayed in, carpets are old and a mess, the catches on window broke, wall paper coming off walls, plaster on ceiling coming off with damp, saniflow units in toilet very noisy and stinks after use, shower curtain too short floor get soaked when having shower, when having shower no room in shower to move and water keeps going up and down, we stayed 2 nights and it cost us £150 I would even pay £20 a night to stay there it a terrible place	0
...	...	...
189532	4\t Always great here. Love the food, staff and short walk to the beach. I would recommend the Oreo Milkshake and Waffle Station Burger	1
189533	4\t Been going here for a while now. All the cakes and savories are clearly handmade, and taste divine! There is a real lack of places like this in surrey, real good quality food, great service, and lovely coffee. Particularly love the cheesecake. Must try	1
189534	4\tThis is a review I meant to do a week ago but it slipped my mind completely which is disgraceful considering the effort the Beasties went to on our behalf. We started as a party of four but realising it was near to my daughters birthday and knowing I wouldn't see her then, we had a traditional &quot;Nearly Birthday&quot;. We asked about a cake as well and more on that later (well not much later as I want to finish this and go home). The ambience is amazing, the preparation of the food incredible, the service just as it should be and it was an absolutely perfect evening. We could have brought our own wine but we drunk theirs and it hit the spot. Back to the cake. I asked for a chocolate cake and expected a sponge covered in chocolate. This was no ordinary cake. I would use the French word but I can't spell it. This was a mountain of chocolate icing hiding a sponge in there somewhere and covered in strawberries and blueberries. It made our evening and it was brought out with the candles I supplied and a rendering of &quot;Happy Birthday to You&quot;. All the food was outstanding but with the passage of time and the amount drunk that night besides remembering Pork in the mix somewhere along with the cake I have forgotten the rest. Dulverton is a totally unique place. We used to live near and now we don't, we often come back for a weekend. Where else could you go to a good pub (Woods) and a good Thai Restaurant all with 100 yards of your B and B in a beautiful Exmoor village. Now two nights will no longer be enough as you have to fit the Exmoor Beastro in too. So thank you guys for a perfect evening and you deserve all the praise and success you are getting.	1
189535	4\tWe were really pleased to receive Sunday lunch as a Christmas present, we arrived at the venue as a party of four, tractor and trailer arrived and we were transported down to the dining area, first stop was at the Yurt for a warming aperitif, quickly warmed around the fire with a friendly introduction to the days events, Although an exceptional cold day we were made to feel very warm with adequate heating, great experience, very friendly staff, brilliant food, wouldn't have missed it for the world, would recommend to anyone!	1
189536	4\t excellent food, Thursday night so not too busy, walked in and seated 7 of us straight away. Simply lovely food and clean and today restaurant. As a seasoned eater of Indian food I class myself as an expert in the cuisine, as such take it from me you won't be disappointed.	1

Figure 3.2: Sample of TripAdvisor Dataset After Conversion to CSV

3.3 Gender Classification Methodology

The proposed Gender Classification methodology of author’s text is described in detail in this section. The basic idea is to leverage the linguistic textual features of text by extracting reference features that assist in computing the count of the most similar tokens in each written text to build the feature vectors that are utilized to improve the accuracy of the proposed methodology. It is worth noting that in the CNN model, the structures of CNN are followed that proceeded in [26], but through experiments different parameters.

Three models comprise the proposed methodology. The first model is applied by three machine-learning classifiers, which are RF, LR and SVM to obtain the prediction probabilities of gender labels. The second model is applied through the pre-used successful structures of the CNN models to obtain gender label probabilities. Both the models are described in the following subsections. Finally, a linear combination mechanism (the third model) is used by combining additional weights with the label probability results of the previous two models to compute the final prediction probability and get the final output as class label (male or female). Figure 3.3 presents the block diagram of the proposed methodology.

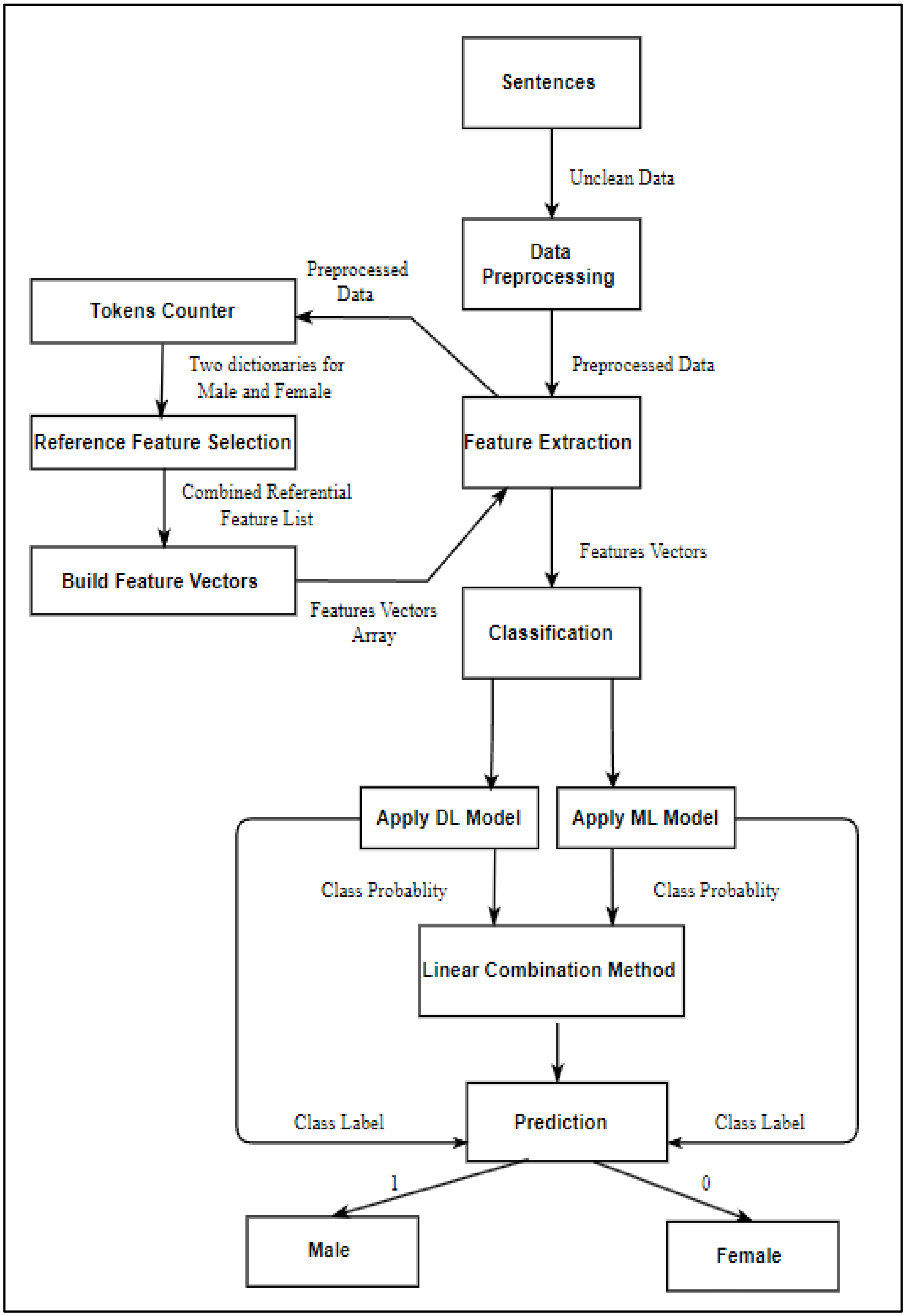


Figure 3.3: Block Diagram of the Proposed Methodology for Gender Classification in Author's Text

The stages of the proposed methodology are as follows:

3.3.1 Data Preprocessing

Data preprocessing must be performed to improve the accuracy of the models. Several preprocessing techniques are presented in the following:

1. Binarization process is performed, meaning that every row of a dataset that is not labeled as female or male is filtered out, and the label column is converted to binary values (i.e., female → 0 and male → 1).
2. Some of common cleaning processes on text are performed which are lowercase, stop-words and punctuation marks removal, tokenization and lemmatization, etc. Each step of them is discussed below.
   - *Lowercase:* is the converting of all texts to lowercase ensures consistency of the dataset.
   - *Stop-words removal*: stop words are most frequent used terms (such as "the," "is," and," for example) that have little to no significance in text.
   - *Punctuation marks removal:* punctuation marks such as commas, periods, and exclamation points are removed to reduce noise in data.
   - *Tokenization:* the process of splitting a text into smaller parts, often tokens, which are words or phrases. This step is important to process and analyze the text.
   - *Lemmatization:* the process of normalizing words while decreasing their vocabulary size by returning them to their basic form, such as "working → work”.

3. A number of additional cleaning processes are performed,
    - Deleting digits.
    - Remove short words (less than two characters).
    - Excluding HTTP links (URLs), HTML tags, hashtags, usernames.
    - Substitute accented characters such as (á → a, ç → c, ü → u and ö → o) etc.
    - Expanding contracted words. Expands words like (I'll → I will) for better text classification
    - Removal duplicate rows.
    - Remove strip extra and trailing spaces.
4. Any row that results in empty text after applying the previous step is excluded to avoid any failure in the subsequent feature vector calculations.

Preprocessing stage is necessary to analyze the text data and then build feature vectors from them.

3.3.2 Feature Extraction

This thesis proposes a new method for feature extraction. The preprocessed data from the previous stage is the input for this stage, while the feature vectors array is its output. Feature extraction stage includes three stages tokens counter, reference feature selection and build feature vectors.

3.3.2.1 Tokens Counter

After preprocessing the textual data in the previous stage, comes the turn of this step is performed, which involves counting the number of tokens within the set of feminine and masculine texts. This means that the inputs are all the aforementioned texts, and then, using two software components. It includes two stages:

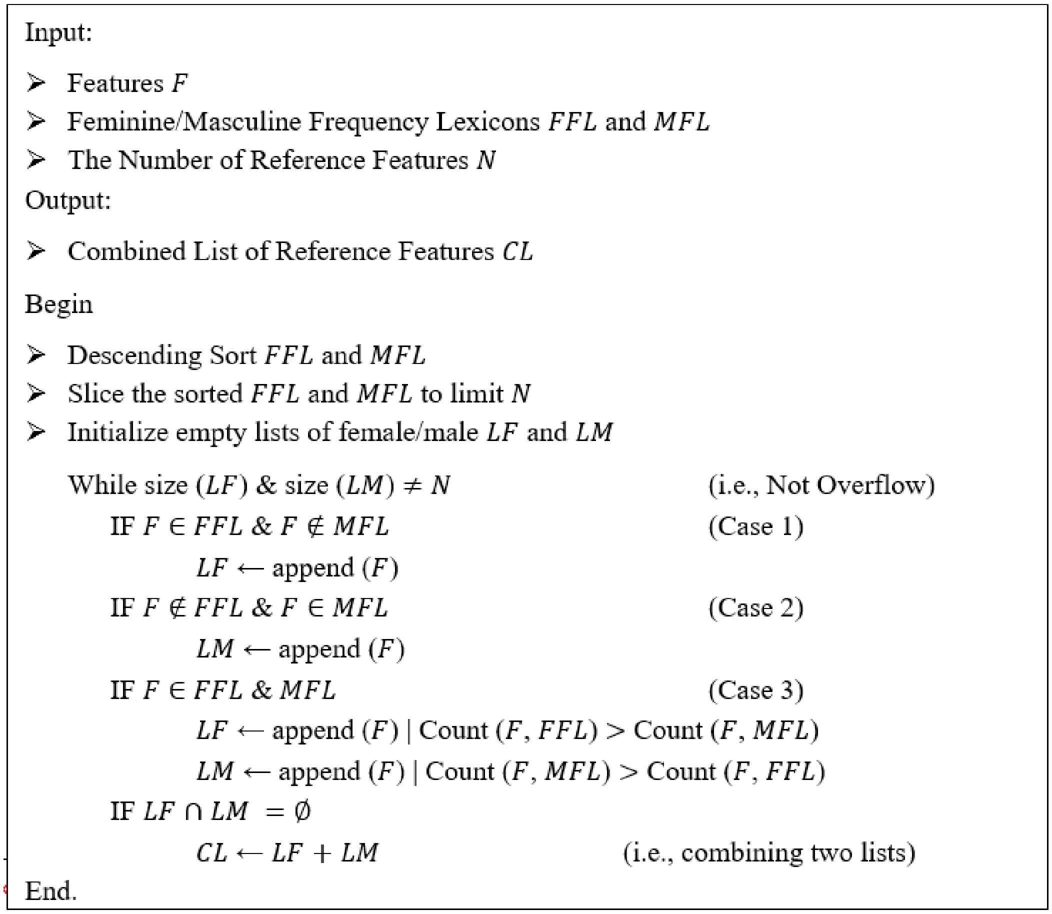
- First Stage: digital vectors of words from spaCy library are used.
- Second Stage: by using Counter Package statistics are performed for each token within the two groups (female/male) to count the number of tokens (token is 1-gram word).

The procedure in this step is repeated once for each group of feminine and masculine texts to produce two dictionaries that play a significant role in the next steps of the proposed methodology.

3.3.2.2 Reference Feature Selection

Reference feature selection is very important because it determines the discriminatory features that distinguish between the two labels (female and male). The outputs obtained in the previous step (tokens counter) are the inputs to this step with an additional parameter (N) that represents the number of reference features for each lexicon. Initially, multiple N (2×N) tokens are deducted from the two frequency dictionaries and then sorted in descending order based on the number of occurrences within each dictionary. Referring to Algorithm 3.1, three cases are used to identify the most frequent reference features within the feminine and masculine frequency lexicons.

Algorithm 3.1: Reference Feature Selection



```
Input:
➢ Features $F$
➢ Feminine/Masculine Frequency Lexicons $FFL$ and $MFL$
➢ The Number of Reference Features $N$
Output:
➢ Combined List of Reference Features $CL$
Begin
➢ Descending Sort $FFL$ and $MFL$
➢ Slice the sorted $FFL$ and $MFL$ to limit $N$
➢ Initialize empty lists of female/male $LF$ and $LM$
    While size $(LF)$ & size $(LM) \neq N$                    (i.e., Not Overflow)
        IF $F \in FFL$ & $F \notin MFL$                       (Case 1)
                $LF \leftarrow$ append $(F)$
        IF $F \notin FFL$ & $F \in MFL$                       (Case 2)
                $LM \leftarrow$ append $(F)$
        IF $F \in FFL$ & $MFL$                                (Case 3)
                $LF \leftarrow$ append $(F)$ | Count $(F, FFL) >$ Count $(F, MFL)$
                $LM \leftarrow$ append $(F)$ | Count $(F, MFL) >$ Count $(F, FFL)$
        IF $LF \cap LM = \emptyset$
                $CL \leftarrow LF + LM$               (i.e., combining two lists)
End.
```

First, if and only if the feature (F) is present in the feminine frequency lexicon, it is added to the list of feminine referential features. Second, the same occurs for the list of masculine referential features only when the feature is present within the masculine frequency lexicon. Finally, in the third case, when the current feature was present in both dictionaries, a comparison was made based on the greater number of occurrences. In other words, if this feature is more frequent in the feminine frequency lexicon than in the masculine lexicon, then it is added to the list of feminine referential features and vice versa. In all cases, the intersection of two lists must be an empty set. This step produces two lists of feminine and masculine referential features, representing the most frequent features within the two related frequency lexicons, which are combined by append masculine referential features to feminine referential features to create the final combined referential features list that have a primary role in the feature vectors construction step.

3.3.2.3 Building Feature Vectors

This step is the core of the proposed methodology. After implementing the previous steps, the outputs are pre-processed text from (preprocessing stage), feminine/masculine frequency lexicons from (tokens counter stage), and a combined list of referential features from (reference feature selection stage). All of these data will be available for processing within the feature vector building step (distinct feature matrix), which takes place in two stages.

- **First Stage**

Calculates the similarity degree using the cosine similarity measure of the tokens' spaCy vectors by comparing the text features (tokens) with each reference feature from the combined list. The similarity score ensures the selection of the most influential and like-context text features (tokens) for the current reference feature.

- **Second Stage**

The Most Similar Feature (MSF) is determined by calculating the maximum similarity degree for each reference feature. In addition, the frequency of MSF is recalled from the two frequency lexicons to compute the MSF overall frequency average after calculating its total frequency (summation of its feminine/masculine frequency), as stated in Eqs. (3.1) and (3.2).

$$MSF = Max\big(Cosine_{sim}(F, RF)\big) \tag{3.1}$$

$$Fvec_{idx} = \begin{cases} MSF_{freq1} \,/\, \big(MSF_{freq1} + MSF_{freq2}\big), & idx < N \\ MSF_{freq2} \,/\, \big(MSF_{freq1} + MSF_{freq2}\big), & idx \geq N \end{cases} \tag{3.2}$$

Where MSF refers to the most similar feature. F and RF are the text features (tokens) and current reference features, respectively. In addition, $Fvec_{idx}$ represents the current location of the feature vector. MSF_{freq1} are the frequency of MSF from masculine frequency lexicon, while MSF_{freq2} the frequency of MSF from feminine frequency lexicon, and their summation refers to the total frequency of MSF. Finally, N represents the size of the non-combined list of reference features (half the size of the feature vector). Due to combined referential features list with 2N size (the first half is masculine and the second half is feminine referential features) and based on index, if idx smaller than N (it means that $Fvec_{idx}$ is for male). While if idx is larger than or equals to N (it means that $Fvec_{idx}$ is for female).

The above equation is applied to all texts of the training set to build a matrix of distinct feature vectors for each text, where vector size is equal to 2×N and its values in [0,1] scale. At the end of this step, a distinct features matrix is built that will represent the cornerstone of gender classification using ML and DL models in parallel in the next step of the proposed methodology.

3.3.3 Applying Classification Models

Two diverse models are used for gender classification in this step through the parallel execution of both ML classifiers and DL model on the matrix of distinct features obtained from the previous steps. The details of each model are described separately by clarifying the types of classifiers or deep structures preferred and their hyperparameter-setting strategy.

a) Machine Learning Model

In this model, three different classifiers, among the best machine - learning classifiers, are tested to confirm the generality of the proposed methodology in terms of the diversity of implementation strategies. RF, SVM, and LR are the classifiers chosen to improve the accuracy of gender classification. All of these classifiers are applied to the distinct feature matrix as input to the model and the output is label class (male or female) with class probability. Table 3.2 shows the parameters that are used for fitting each classifier, where many experiments are conducted to achieve the best classification accuracy by harmonizing these parameters together. Later, in the last step of the proposed methodology, the results of this model and the next model (DL) are merged using an innovative LC method.

Table 3.2: Parameters of ML Model

Classifier	Parameters Names
RF	Number of Estimators, Criterion, Maximum Depth, Minimum Samples Split, Minimum Samples Leaf, Maximum Features, Bootstrap
SVM	Kernel, C, Gamma, Class Weight, Probability
LR	Penalty, C, Solver, Maximum Iteration, Class Weight

The important parameters of the RF classifier are listed in Table 3.2. Many experiments have been conducted to get the final parameters of the classifiers RF, LR and SVM that have obtained the best accuracy results as shown in the following.

- *Random Forest*

The number of trees in the forest is adjusted to balance decreasing variability and performance by setting Number of Estimators to 100. By using the Criterion=’Gini’, the split quality is evaluated which improves classification accuracy. The value Maximum Depth=10 is set to avoid overfitting and restrict complexity. Overfitting is avoided when Minimum Samples Leaf is 2, and splits are triggered when Minimum Samples Split is 2. Model robustness is strengthened by setting Bootstrap=True and selecting Maximum Features='sqrt' to allow for tree variation.

- *Support Vector Machine*

SVM performance is influenced by the using of parameters like with, C=1.0 which is reducing testing and training errors overfitting. The linear kernel coefficient, set by Gamma='scale', is adjusted based on the number of features in the data. This influences the complexity of the decision boundary. In addition, setting the Class Weight to ‘balance’ and Probability to ‘True’ can significantly improve the model's performance.

- *Logistic Regression*

The prediction quality in the LR model is significantly influenced by each parameter, and large weights are minimized using regularization (Penalty='l2') to prevent overfitting. Model performance and complexity are balanced by C=1.0. The Solver liblinear is great for binary classification. The Maximum Iteration=100 is set to ensure efficient iteration termination. To ensure fair class representation, Class Weight='balanced' is used.

Finally, the results are guaranteed across the runs using random_state=42. Adjusting these parameters can help reduce overfitting, manage complexity, and improve accuracy.

b) Deep Learning Model

Many structures have been tested in deep-learning model, as shown in Table 3.3. The CNN model is chosen to implement on a distinct feature matrix as the input to the model and the output is label class (male or female) with class probability. In addition, more than one CNN structure is tested to confirm the generalizability of the proposed methodology to diverse procedures, not just the diversity of data. The compilation procedure for every CNN model starts by assigning parameters that are supplied to the model's fitting operation. Such parameters include the type of classification, activation function, loss function, and optimizer type.

The Sparse Categorical Cross Entropy loss function, the Adam optimizer with well selected hyperparameters, and sequential layers with ReLU and Softmax activation functions are the components of this classification model. The layers are: MaxPooling1D for down sampling the input, Dropout for avoiding overfitting, Dense layers for feedforward NN operations, and Conv1D for 1-dimensional convolutional filtering. The loss function and the Adam optimizer is used to assemble the model with a learning rate of 0.001. The model is fit to the training data across 20 epochs with a batch size of 32 during training, and then the model's performance is assessed using the validation data. This configuration maximizes our model's capacity for learning and producing precise forecasts while avoiding overfitting or complexity.

Different structural CNN models (CNN$_1$, CNN$_2$, and CNN$_3$) were applied by adapting several hyperparameters shown in the Table 3.3 to achieve superior accuracy within gender classification. Each of layers in the Table below are discussed in section (2.8.1).

Table 3.3: Structure Layers of DL Model

Model	**Structure Layers**
CNN$_1$	Conv + Pool + Dropout + FC
CNN$_2$	Conv + Pool + Conv + Pool + Dropout + FC
CNN$_3$	Conv + Pool + Conv + FC + Dropout + FC

3.3.4 Linear Combination Method

By the end of the classification process, every record is assigned two probabilities: feminine and masculine class probabilities. Depending on these probabilities and the Opposite Weight (OW), a Linear Combination (LC) method (as mentioned in section 2.12) can be used to predict the related final probability (as referred in Algorithm 3.2). Therefore, two stages are considered in this study.

- **First Stage**

Feminine and masculine class probabilities of the first model (Machine Learning Classifier) are collected in a list of tuples in order to organize them for the LC operation later, as in Eq. (3.3).

$$CLS_{probs} = \left[\left(\left(F_{prob}, M_{prob}\right) \middle| \forall\, F_{prob}, M_{prob} \in CLS_{preds}\right)\right] \tag{3.3}$$

Where CLS_{probs} refers to the list of tuples female/male probabilities of a classifier. F_{prob}, M_{prob} are probabilities corresponding to each class label, female and male, respectively. In addition, CLS_{preds} represent the probability predictions of each classifier.

- **Second Stage**

The feminine and masculine class probabilities of the second model (Deep Learning Model) are similarly assembled in a list of tuples so as to arrange them for the LC method next. The assembly formula as in Eq. (3.4):

$$MOL_{probs} = \left[\left((F_{prob}, M_{prob}) \,\middle|\, \forall\, F_{prob}, M_{prob} \in MOL_{preds}\right)\right] \tag{3.4}$$

Where MOL_{probs} refers to the list of tuples female/male probabilities of DL model. F_{prob}, M_{prob} are class label probabilities for female and male, respectively. In addition, MOL_{preds} represent the probability predictions of each model structure.

Therefore, by retrieving these probabilities using Eqs. (3.3) and (3.4), the LC method can be utilized by employing an additional OW. Then, it linearly incorporates the probabilities of the two models, with integrating OW ranging from 0.1 to 0.9. Thus, LC mechanism has been observed to set the appropriate OW that will provide the finest possible classify, as in the following Eqs (3.5), (3.6) and (3.7).

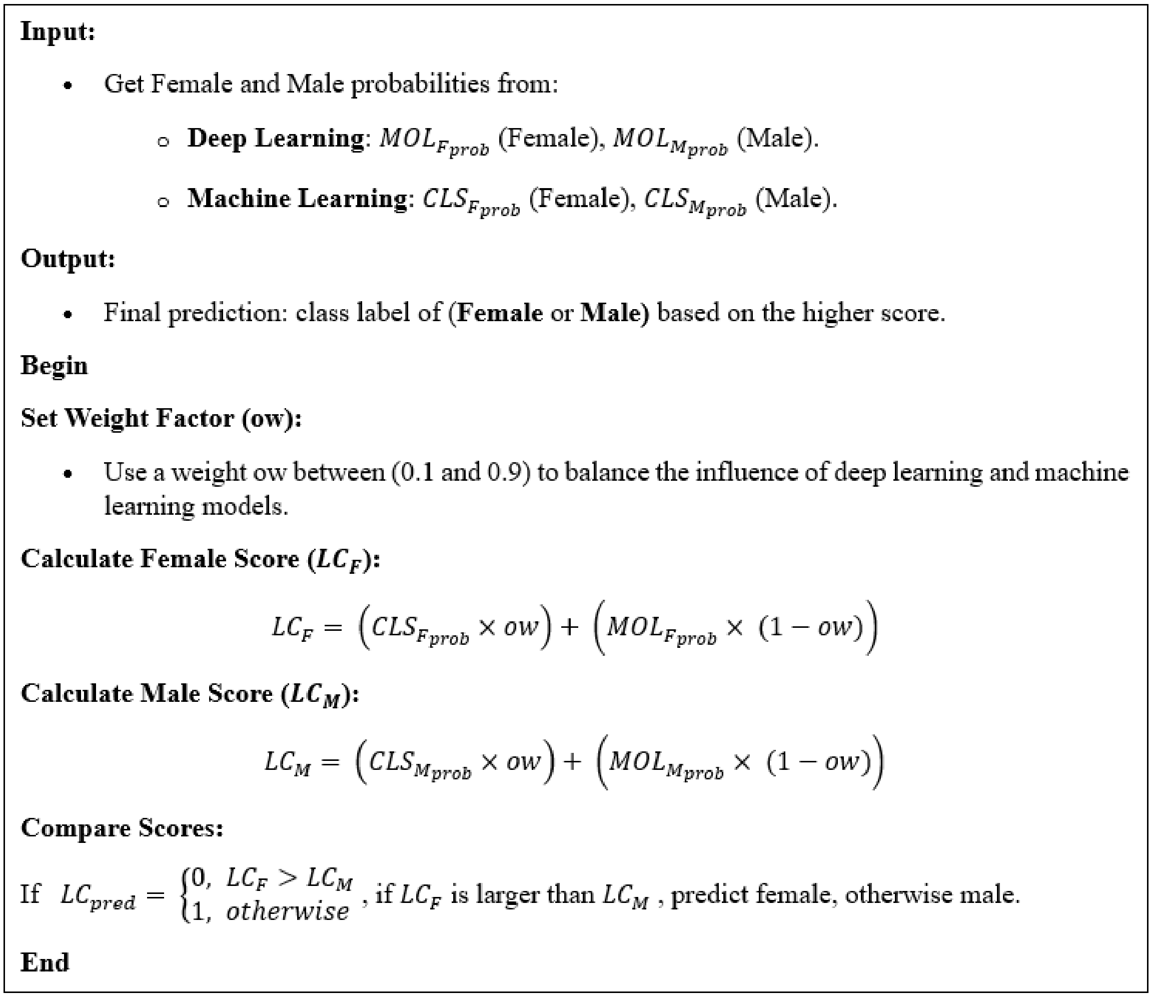
$$LC_F = \left(CLS_{F_{prob}} \times ow\right) + \left(MOL_{F_{prob}} \times (1 - ow)\right), \; \forall\, ow \in [0.1,0.9] \tag{3.5}$$

$$LC_M = \left(CLS_{M_{prob}} \times ow\right) + \left(MOL_{M_{prob}} \times (1 - ow)\right), \; \forall\, ow \in [0.1,0.9] \tag{3.6}$$

$$LC_{pred} = \begin{cases} 0, & LC_F > LC_M \\ 1, & otherwise \end{cases} \tag{3.7}$$

Where LC_F, LC_M are the combined probabilities of female/male labels respectively. $CLS_{F_{prob}}, CLS_{M_{prob}}$ are the classifier probability predictions related to each class label (female and male), respectively. Additionally, $MOL_{F_{prob}}, MOL_{M_{prob}}$ are the DL model probability predictions related to each class label respectively. Lastly, LC_{pred} refers to the final class prediction. Notably, OW should not be 1 or 0 according to the above equation regardless of other values due to if OW = 1 then LC probability prediction will be exactly as one model and completely discard the other one and vice versa when OW = 0.

Algorithm 3.2: Linear Combination Method



```
Input:
    • Get Female and Male probabilities from:
        ○ Deep Learning: $MOL_{F_{prob}}$ (Female), $MOL_{M_{prob}}$ (Male).
        ○ Machine Learning: $CLS_{F_{prob}}$ (Female), $CLS_{M_{prob}}$ (Male).
Output:
    • Final prediction: class label of (Female or Male) based on the higher score.
Begin
Set Weight Factor (ow):
    • Use a weight ow between (0.1 and 0.9) to balance the influence of deep learning and machine
      learning models.
Calculate Female Score ($LC_F$):
        $LC_F = \left(CLS_{F_{prob}} \times ow\right) + \left(MOL_{F_{prob}} \times (1 - ow)\right)$
Calculate Male Score ($LC_M$):
        $LC_M = \left(CLS_{M_{prob}} \times ow\right) + \left(MOL_{M_{prob}} \times (1 - ow)\right)$
Compare Scores:
If $LC_{pred} = \begin{cases} 0, & LC_F > LC_M \\ 1, & otherwise \end{cases}$ , if $LC_F$ is larger than $LC_M$ , predict female, otherwise male.
End
```

3.3.5 Performance Evaluation

In this step, the proposed system is evaluated using a testing dataset to assess its accuracy. To ensure the reliability and precision of the results, five-fold cross-validation is employed. Accuracy is utilized as the sole evaluation metric to measure the system's performance. The training and testing process is repeated five times, with each iteration involving the division of the dataset into five distinct folds. Cross-validation serves as a robust method for evaluating model performance on unseen data by resampling under a defined parameter, referred to as 'K'. This approach provides an estimate of how the model is expected to perform on new datasets. In this study, a five-fold cross-validation is applied, where the dataset is systematically partitioned into training and testing sets. The final performance result is derived by calculating the average accuracy across all folds, offering a comprehensive evaluation of the model's effectiveness.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Overview

Results discussion of extensive experiments will be displayed in this chapter for the three models includes machine learning model, deep learning model and linear combination method. In this chapter the experiments have been performed using the proposed methodology to ascertain its results on two real-world datasets, namely Twitter and TripAdvisor. The accuracy metric has been employed to assess the classification outcomes of the proposed method against the methods in other studies. To remember, the evaluation has been driven through the process of five-fold cross-validation. The experimental process and detailed findings are illustrated in the following sections.

4.2 System Requirements

In order to carry out machine learning and deep learning on a dataset, the software or program needs to run on a computer system with adequate computational resources. Therefore, the proposed system is set up as follows:

Processor: AMD Ryzen 7 3700U with Radeon Vega Mobile Gfx 2.30 GHz
RAM: 8.00 GB of RAM (6.94 GB useable) computer
Hard Disk: 512 GB SSD
Operating System: Windows 10 – 64bit
Programming Environment: Google Collaborator
Programming Language: Python

4.3 The Results of Machine Learning Model

In this section, the classification results of the ML model have been presented. As previously mentioned, this model has been tested using three common machine learning classifiers, namely RF, LR, and SVM. Its experiments have been carried out on Twitter and TripAdvisor datasets. Due to the presence of additional description text in Twitter, the experiments of five folds have been implemented on textual comments, descriptions, and both as shown in Table 4.1. That is, each classifier has been trained and tested first on the descriptive text, then the textual comments, and yet on the combined text of both together, as shown in Figure 4.1 (Twitter-D, Twitter-T, Twitter-TD, and TripAdvisor -T), respectively.

Table 4.1: Classification Accuracy Results of ML Model on Twitter with N=120

Classifier	**Target Data**	**Accuracy (%)**					
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average of Folds
RF	Description	**88.7**	88.2	**88.7**	87.2	86.4	87.84
	Text	84.2	84.4	**84.7**	83.5	84.4	84.24
	Description + Text	84.7	86.1	**86.9**	85.6	85.5	85.76
LR	Description	**87.6**	87.3	87.2	87	86.4	87.1
	Text	83.1	84	**84.2**	83.2	83.7	83.64
	Description + Text	85.6	86.3	**86.6**	86.4	86	86.18
SVM	Description	**88.1**	87.1	87	86.5	85.9	86.92
	Text	83.4	83.5	**83.8**	82.9	83.6	83.44
	Description + Text	85.9	85.9	**86.7**	85.9	85.3	85.94

On TripAdvisor dataset, the three classifiers have been only implemented on textual comments, as shown in Table 4.2 and Figure 4.1. It is worth noting that all of these results have been recorded when the number of referential features (N) are equal to 120 (the combined vector size of each record comprises 240 referential features), which at this point the best results have captured.

Table 4.2: Classification Accuracy Results of ML Model on TripAdvisor with N=120

Classifier	**Target Data**	**Accuracy (%)**					
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average of Folds
RF	Text	**71.4**	71.1	71.6	70.7	70.2	71
LR	Text	**76.3**	74.8	75.5	74.6	74.1	75.06
SVM	Text	**75.7**	74.3	75.6	74.6	74	74.84

From the results in Tables 4.1 and 4.2, folds 1 and 3 have been offered the best outcomes as shown in bold. Figure 4.1 shows the results of average accuracy of 5-folds by using ML model classifiers on all texts of both datasets. The best accuracy obtained was 87.8% in Twitter-D by using RF. And the best accuracy in TripAdvisor was 75.1% by using LR.

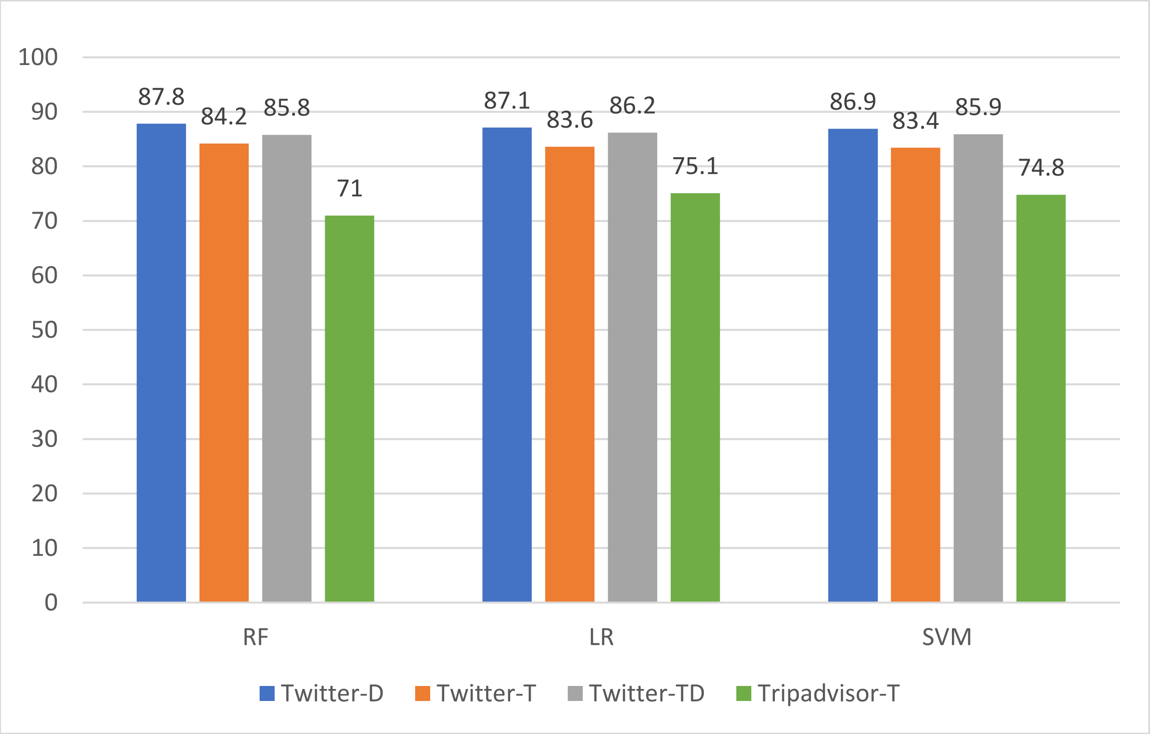


Figure 4.1: Results of Average Accuracy using ML Model on both Datasets

In the previous tables, the results showed that the ML model based on each classifier has been achieved clear generality in terms of the convergence results within each dataset. The three classifiers achieved great convergence in their results on Twitter (87.6%, 88.1%, and 88.7%) on descriptive text because it has more unique referential features than textual comments. In particular, RF classifier was the best of them on all text data by a slight margin by using 5-folds results, as shown in Table 4.1.

Notably, a decline in accuracy to about 83% is noticed by all classifiers when tested on textual comments. Two reasons for such case. The first one confirms that descriptive text is distinguished from textual comments with greater uniqueness of the referential features. Additionally, the second reason is due to the average number of tokens, which is much larger in comments compared to descriptive text. The following case confirms the above interpretation which is, when the descriptions and comments are combined, the accuracy increased within Twitter from 84.7% to 86.9%. This is better than when the classifiers implement only on the text.

Additionally, the matter is not much different within TripAdvisor, where the accuracy values are close on almost all classifiers. Particularly, LR classifier with 76.3% was the best among them on the available textual comments. It is worth noting that the average number of tokens of TripAdvisor is much larger than Twitter (see Table 3.1). This huge average might cause the unwanted token to be the MSF to the referential features. Because of this the accuracy values on TripAdvisor have appeared much less than the values on Twitter (refer to Figure 4.1).

4.4 The Results of Deep Learning Model

The DL model has been examined by measuring its classification outcomes. As previously said, this model has been tested using three structure mechanisms which are called CNN$_1$, CNN$_2$, and CNN$_3$ (See Table 3.3). As before, the experimentation has been carried out on both Twitter and TripAdvisor datasets.

On Twitter, each CNN model has been trained and tested on description, text comments, and combined version of both texts, as shown in Figure 4.2 (i.e., Twitter-D, Twitter-T, Twitter-TD, and TripAdvisor -T), respectively.

Table 4.3: Classification Accuracy Results of DL Model on Twitter with N = 120

Classifier	**Target Data**	**Accuracy (%)**					
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average of Folds
CNN$_1$	Description	**90.3**	88.3	89.8	88.6	87.4	88.88
	Text	84.5	85	84	84.1	**85.5**	84.62
	Description + Text	88.2	**89.5**	89	89.4	89.3	89.08
CNN$_2$	Description	88.5	88.2	**88.9**	88.1	87.5	88.24
	Text	83.5	**85.8**	84.7	84.2	84.6	84.56
	Description + Text	87.6	88.6	**89.4**	87.8	89.1	88.5
CNN$_3$	Description	88.1	88.2	**89.2**	87.7	86.2	87.88
	Text	83.7	83.6	83.7	82.1	**84.3**	83.48
	Description + Text	88	86.6	**88.5**	88.1	87	87.64

These CNN models have been executed on available textual data of TripAdvisor, as shown in Table 4.4. Similarly, the finest outcomes of this approach have been captured when the number of referential features (N) are equal to 120.

Table 4.4: Classification Accuracy Results of DL Model on TripAdvisor with N = 120

Classifier	**Target Data**	**Accuracy (%)**					
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average of Folds
CNN$_1$	Text	76.6	**76.7**	**76.7**	76.2	75.5	76.34
CNN$_2$	Text	**77**	75.2	75.4	75	76.3	75.78
CNN$_3$	Text	73.9	72.3	74.9	74.1	**75.2**	74.08

From the above table, except for fold 4, all other folds have been offered the best outcomes in terms of the three CNN models as shown in bold. Figure 4.2 shows the average results of 5-folds in terms of all models by using DL model on all texts of both datasets. The best accuracies obtained in Twitter-TD, TripAdvisor by using CNN$_1$ were (89.1%,76.3%), respectively.

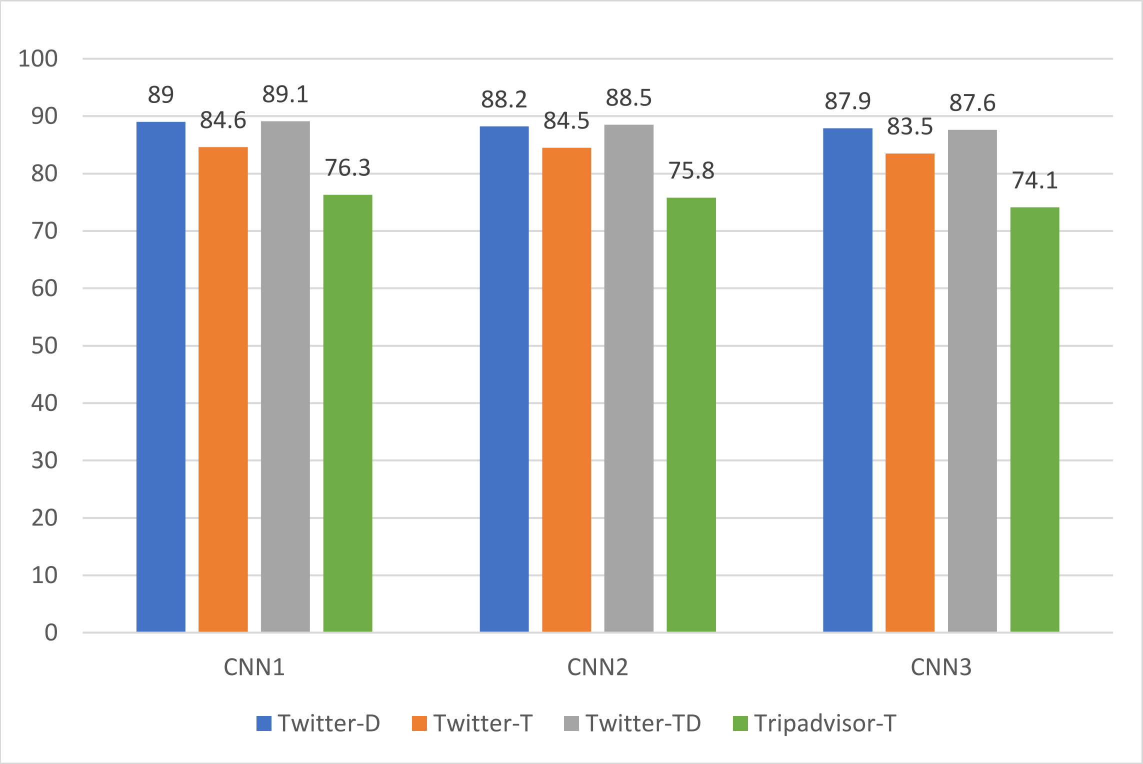


Figure 4.2: Results of Average Accuracy using DL Model on both Datasets

As shown in the previous two tables, the deep learning model has been accomplished obvious generality in terms of the convergence outcomes within each dataset. The CNN mechanisms have reached remarkable accuracy (about 89%) on Description/Text (D/T) data of Twitter because they combine more unique referential features than the individual descriptive or textual comments. In particular, CNN_1 model was the best of them by an obvious margin, as shown in Table 4.3.

Notably, the accuracy values decreased to about 82-85% in all models when tested on text comments only. This is due to two reasons. First, it ensures that combined D/T is differentiated from just textual comments with more degree of distinctiveness of the referential features. Second, its due to the large number of tokens has less negative influence on accuracy compared to only text without description.

On TripAdvisor, the case is much similar, except that the maximum value of accuracy has been reached up to about 77%. Particularly, the accuracy is close in almost all models, and CNN_1 with 76.7% was the most satisfactory among them on the available textual comments. The reason, that the classification accuracy restricted to under 80% (as a threshold) compared to the accuracies achieved on Twitter, is that the average number of tokens on TripAdvisor is double the number on Twitter (see Table 3.1). This extremist difference could drive the undesirable token to be picked over another token as the MSF to the referential features. Thus, the accuracy values on TripAdvisor have lesser than the ones outlined on Twitter texts (refer to Figure 4.2).

4.5 The Results of Linear Combination Method

This step of the final classification is to compute and combine the results of the above two models. These depended on a bunch of common classifiers and multiple CNN models. Their class probabilities, however, are separately collected from Eqs. (3.3) and (3.4) respectively. Then the combination process has been accomplished using a critical factor (viz., Opposite Weight) as previously referred to in Eqs. (3.5), (3.6) and (3.7). Essentially, the idea concerns the examination of the impact of combining feminine/masculine class probabilities of two diverse models on the accuracy of Gender Classification. According to this vision, the Opposite Weight values are shifted from 0.1 to 0.9 with an increment of (0.1). The best results from deep learning were combined by using CNN$_1$ structure which obtained against other structure (CNN$_2$, CNN$_3$) with the best fold from machine learning classifiers on both datasets to get the finest results for the linear combination method, as shown in Tables 4.5 and 4.6.

Table 4.5: Linear Combination Accuracy Results of (best fold of each RF, LR, and SVM with CNN$_1$) on Twitter

Target Data	**Fold**	**OW** **LC**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Description	1	RF + CNN$_1$	90.3	90.4	90.6	90.6	90.7	**90.8**	90.5	90.1	89.6
Text	5		85.5	85.8	86	86	86	**86.1**	**86.1**	85.6	85.3
Description + Text	2		89.6	89.8	89.8	**89.9**	89.7	89.4	88.9	88.5	88.1
Description	1	LR + CNN$_1$	90.4	90.6	**90.7**	90.6	90.4	90.3	89.6	89	88.5
Text	5		85.4	85.5	85.8	85.9	**86.2**	85.9	85.4	84.6	84.1
Description + Text	4		89.7	**89.9**	89.7	89.8	89.8	89.2	88.5	87.7	87
Description	1	SVM + CNN$_1$	90.4	90.6	**90.7**	90.6	90.4	90.2	89.7	89.3	88.5
Text	2		85.2	85.7	85.9	86	**86.2**	86	85.4	85.1	84.4
Description + Text	4		89.6	89.9	89.9	89.8	**89.9**	89.3	88.4	87.7	86.7

Table 4.6: Linear Combination Accuracy Results of (best fold of each RF, LR, and SVM with CNN_1) on TripAdvisor

Target Data	**Fold**	**W** **LC**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Text	3	RF + CNN_1	76.9	76.9	77	77.2	**77.4**	77.1	76.8	76	74.1
Text	3	LR + CNN_1	76.9	77	76.9	77.2	**77.4**	**77.4**	77.1	76.8	76
Text	3	SVM + CNN_1	76.9	77.1	77	77.2	**77.3**	77.2	77.2	76.6	76.2

Based on the linear combination experiments, it has been deduced that the optimum value of the OW is 0.3, 0.4, 0.5, 0.6 on Twitter and 0.5, 0.6 on TripAdvisor. This means the more likely this weight to be evenly distributed over the two models the more accurate the outcomes and vice versa. In this case, the ultimate classify outcome is largely depend on the probabilities of the constructional models (refer to Eqs. (3.5), (3.6) and (3.7)).

The last classification results follow the same routine when applied on both datasets. In terms of accuracy measure, the most satisfactory outcome has been captured when the OW value is in the mid-range (viz., 0.3 to 0.6), as the accuracy values shown in Figures 4.3 and 4.4. These mid-range weight values signify that both models Female/Male class probabilities have an influential role in computing the final classify value. This point is logical, given that they both contribute with convergent rates of weight, especially when their original accuracies are somewhat close. So, this explains the high accuracy of the combination process, as both models have relatively equally contributed to the share.

The cases of inability to produce new accuracy of such process are occurred when the weight equals 1 or 0, owing to mathematical fact. In other words, if the first part of the equation is multiplied by the weight value of 1 then the other will surely multiply by zero. Therefore, these values are discarded from the interval range refer to the last parts of Eqs. (3.5) and (3.6). This is the reason to rely only on the range from 0.1 to 0.9. Thus, it instructs a boost in the classification accuracy, which reflects thoroughly on the overall performance.

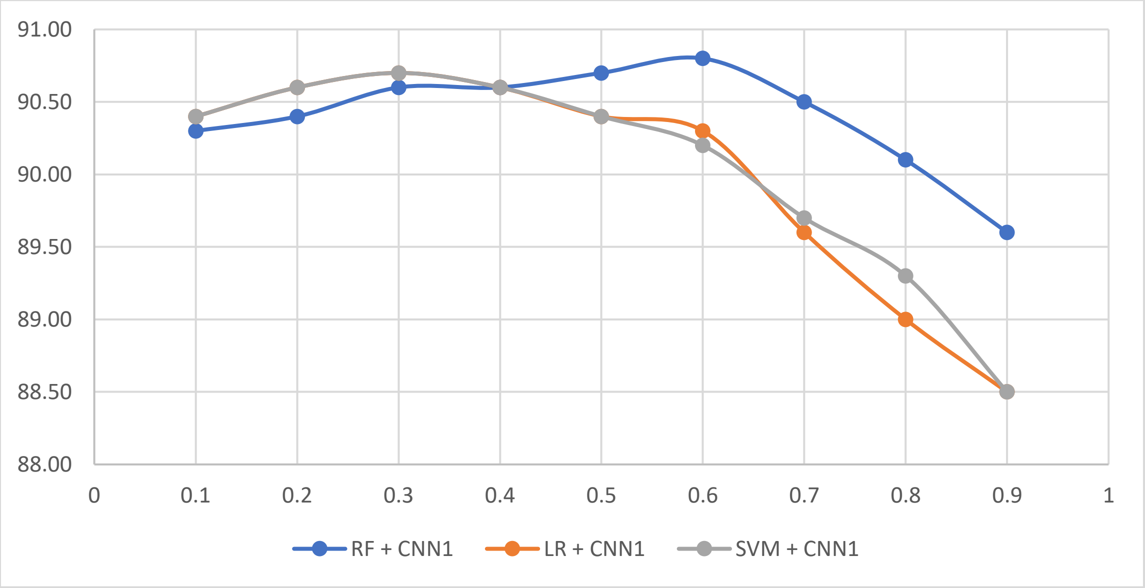


Figure 4.3: Best Combination Accuracy on Twitter

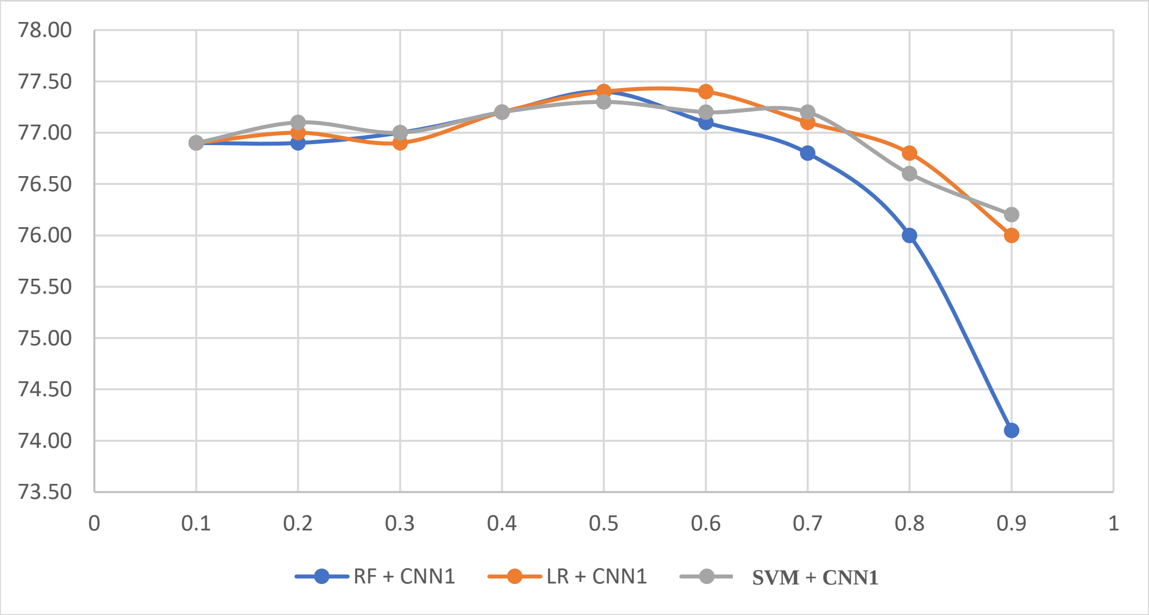


Figure 4.4: Best Combination Accuracy on TripAdvisor

Certainly, merging feminine/masculine class probabilities of two diversity models by linearly fusing them utilizing the optimum value of OW could raise the accuracy of the final classify value. This improvement is directly related to how we develop the equations (3.5), (3.6) and (3.7). According to Figures. 4.3 and 4.4, its noted that the qualitative increase on both datasets is in the middle of the graphs compared to the edges. This is confirming the impact of combining feminine/masculine class probabilities of two diverse models on the accuracy of Gender Classification. Accordingly, the classes of many mislabeled examples have been corrected, thus improving the overall classification accuracy.

4.6 Machine Learning Results Comparison with Other Studies

For each one of the formerly mentioned comparison methods, their results only available on Twitter are recorded as encountered in their respective articles and as depicted in Table 4.7. For simplicity, we will call the comparison methods as CM_1 and CM_2, where CM_1 has been implemented on just the textual data of Twitter while CM_2 was applied on description, text, and both together, as in the proposed method.

In this section, the proposed method findings of the machine learning model are only demonstrated on Twitter for fair comparison based on the same classifiers (RF, LR, and SVM). For the best of our knowledge, there is no one invested on TripAdvisor dataset in the field of Gender Classification, so the comparison will only be detailed on Twitter dataset.

According to Table 4.7, CM_1 recorded the poorest outcomes on textual comments in terms of the accuracy metric of all classifiers. Besides, CM_2 had slightly better performance on text as reached to maximum value of 63% by LR and 70% by RF on D/T. This explains their lack to choose reference features which are filtered out any undesired features. On the contrary, the proposed method has employed the algorithm of selection of the referential features, which unarguably demonstrates its superiority over CM_1 and CM_2 on any text data of Twitter as illustrated in Table 4.7.

Table 4.7: The Results of Our Method and Comparison Methods on Twitter

Method	**Target Data**	**Classifier**	**Accuracy (%)**
CM_1 [4]	Text	RF	48.46
		LR	**57.14**
		SVM	52.67
CM_2 [32]	Description	RF	65
		LR	**66**
		SVM	65
	Text	RF	59
		LR	**63**
		SVM	61
	Description + Text	RF	**70**
		LR	63
		SVM	61
Our Method (ML Approach)	Description	RF	**87.8**
		LR	87.1
		SVM	86.9
	Text	RF	**84.2**
		LR	83.6
		SVM	83.4
	Description + Text	RF	85.8
		LR	**86.2**
		SVM	85.9

The sovereignty of the proposed method is due to its uniqueness scheme (see Figure 3.1). Starting with feature filtering, similarity matching, and ending with frequency counting. Moreover, the most noticeable reason is fitting the finest N (size of each Female/Male referential features) of the feature vector matrix. This would collect more useful information, thus get rid of useless and redundant details.

For a more reasonable view of the overall improvement that the method of the machine learning model accomplishes, improvement ratios are calculated to be compared to the comparison methods, as illustrated in Table 4.8. These rates have been obtained by subtracting the accuracy values of our method from the accuracy values of other methods (refer to Table 4.7). The more positive the difference between our model and the comparison methods, the more improvements it produced. It is worth noting that the accuracy values on text are used when computing the improvement over CM_1. On CM_2, we selected the best shots (marked with an underscore in Table 4.7) of its accuracies on description and/or text and depend on the respective metric to compute the overall improvement over diverse textual data.

Table 4.8: The Improvements of our Method (ML Model) Compared to other Methods in terms of Accuracy (Twitter)

Metric	**Classifier**	**Comparison Methods**	
		CM_1 [4]	CM_2 [32]
Accuracy (%)	RF	35.74	15.8
	LR	26.46	21.1
	SVM	30.73	21.9

Finally, all result values of this section showed the ability of the proposed method that is based on referential features to provide finer classification marks, as previously revealed in Tables (4.1 - 4.8). Such results support the choice of employing the notion of referential features for a distinctive gender classification system.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1 Conclusions

In the following, the conclusions of the proposed three models are presented together. These are ML model, DL model and a linear combination method of both ML and DL. The conclusions have been drawn from the extensive experiments that were conducted. The main conclusions will be presented based on the results obtained from the corresponding models on related datasets. Following conclusions have been obtained:

1. The step of data preprocessing and adopting the feature counting process has considerably helped in raising the accuracy achieved of gender classification.
2. The algorithm of the reference feature selection, which involves choosing the most similar-countable features, has greatly assisted in leading to more fine results.
3. Adopting the idea of text similarity within the frequences of its outcomes assisted in obtaining better candidate features, and hence a better method performance.
4. Concerning the influence of merging female/male classify probabilities on the accuracy of final prediction, it has found that such a combination contributes positively to increase the final classification accuracy.
5. The brief tweets did not impact the results. In contrast, the general domain of Twitter produced remarkable outcomes due to its diverse range of topics and average token count, making the extracted features more distinctive. On the other hand, the specific domain of TripAdvisor, with twice the average token count of Twitter and a specific topic, led to lower accuracy.

6. Utilizing lemmatization over stemming was crucial to minimize term loss and distortion. Lemmatization retains the original meaning of words, providing more accurate representations. This strategy improved the overall quality and relevance of the results.
7. The most significant contribution to the excellent outcomes and perfect features for differentiating between males and females came from the newly developed and suggested feature extraction method.
8. The gender classification methodology has improved significantly with the use of the linear combination method, which combined the label probabilities of both ML and DL. An LC method of the two distinct classification models with a factor called OW can improve the overall classification accuracy.

5.2 Future Work

- Developing an LSTM-based DL system for gender classification and compare its results to our CNN-based DL approach.
- Implementing and testing the proposed method on different language style by classifying gender based on Arabic-language dataset.
- Extending the test of the proposed gender classification method by crawling text information from various sources articles and blog depending on overall emotion (includes sadness, happiness, fear) and part of speech.
- Building a new gender classification method for the healthcare area based of the comments of patients by utilizing the strategies used in this work.

References

[1] K. Alsmearat, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan, “Author gender identification from Arabic text,” *Journal of Information Security and Applications*, vol. 35, pp. 85–95, Aug. 2017, doi: 10.1016/j.jisa.2017.06.003.

[2] Grosso, Enrico, et al. "Understanding critical factors in appearance-based gender categorization." *Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*. Springer Berlin Heidelberg, 2012. https://doi.org/10.1007/978-3-642-33868-7_28.

[3] W. Xu, Y. Zhuang, X. Long, Y. Wu, and F. Lin, “Human gender classification: a review,” *Int J Biom*, vol. 8, no. 3/4, p. 275, 2016, doi: 10.1504/ijbm.2016.10003589.

[4] P. Vashisth and K. Meehan, “Gender Classification using Twitter Text Data,” in *2020 31st Irish Signals and Systems Conference, ISSC 2020*, Institute of Electrical and Electronics Engineers Inc., Jun. 2020. doi: 10.1109/ISSC49989.2020.9180161.

[5] S. Mukherjee and P. K. Bala, “Gender classification of microblog text based on authorial style,” *Information Systems and e-Business Management*, vol. 15, no. 1, pp. 117–138, Feb. 2017, doi: 10.1007/s10257-016-0312-0.

[6] A. R. Khan *et al.*, “Authentication through gender classification from iris images using support vector machine,” *Microsc Res Tech*, vol. 84, no. 11, pp. 2666–2676, Nov. 2021, doi: 10.1002/jemt.23816.

[7] K. Alsmearat, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan, “Author gender identification from Arabic text,” *Journal of Information Security and Applications*, vol. 35, pp. 85–95, Aug. 2017, doi: 10.1016/j.jisa.2017.06.003.

[8] S. Dargan, M. Kumar, A. Mittal, and K. Kumar, “Handwriting-based gender classification using machine learning techniques,” *Multimed Tools Appl*, 2023, doi: 10.1007/s11042-023-16354-1.

[9] Van Buskirk, Ian, Aaron Clauset, and Daniel B. Larremore. "An open-source cultural consensus approach to name-based gender classification." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 17. 2023.https://doi.org/10.1609/icwsm.v17i1.22195

[10] Safoq, Mais Saad. "Human Gender Prediction by Face Images Based on Convolution Neural Network." *Journal of Kerbala University* 21.1 (2024).

[11] Y. S. TAŞPINAR, M. M. SARITAŞ, İ. ÇINAR, and M. KOKLU, “Gender Determination Using Voice Data,” *International Journal of Applied Mathematics Electronics and Computers*, vol. 8, no. 4, pp. 232–235, Dec. 2020, doi: 10.18100/ijamec.809476.

[12] Owen, Patricia R., and Monica Padron. "The language of toys: Gendered language in toy advertisements." (2015).

[13] M. Thelwall, “Gender bias in sentiment analysis,” *Online Information Review*, vol. 42, no. 1, pp. 45–57, 2018, doi: 10.1108/OIR-05-2017-0139.

[14] E. Teso, M. Olmedilla, M. R. Martínez-Torres, and S. L. Toral, “Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective,” *Technol Forecast Soc Change*, vol. 129, pp. 131–142, Apr. 2018, doi: 10.1016/j.techfore.2017.12.018.

[15] J. C. French *et al.*, “Gender and Letters of Recommendation: A Linguistic Comparison of the Impact of Gender on General Surgery Residency Applicants ☆,” *J Surg Educ*, vol. 76, no. 4, pp. 899–905, Jul. 2019, doi: 10.1016/j.jsurg.2018.12.007.

[16] S. A. Alanazi, “Toward identifying features for automatic gender detection: A corpus creation and analysis,” *IEEE Access*, vol. 7, pp. 111931–111943, 2019, doi: 10.1109/ACCESS.2019.2932026.

[17] L. Balachandra, K. Fischer, and C. Brush, “Do (women’s) words matter? The influence of gendered language in entrepreneurial pitching,” *Journal of Business Venturing Insights*, vol. 15, Jun. 2021, doi: 10.1016/j.jbvi.2021.e00224.

[18] A. Sboev, T. Litvinova, D. Gudovskikh, R. Rybka, and I. Moloshnikov, “Machine Learning Models of Text Categorization by Author Gender Using Topic-independent Features,” in *Procedia Computer Science*, Elsevier B.V., 2016, pp. 135–142. doi: 10.1016/j.procs.2016.11.017.

[19] J. A. B. L. Filho, R. Pasti, and L. N. De Castro, “Gender classification of twitter data based on textual meta-attributes extraction,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2016, pp. 1025–1034. doi: 10.1007/978-3-319-31232-3_97.

[20] E. Alsukhni and Q. Alequr, “Investigating the Use of Machine Learning Algorithms in Detecting Gender of the Arabic Tweet Author,” 2016. [Online]. Available: www.ijacsa.thesai.org

[21] S. Mukherjee and P. K. Bala, “Gender classification of microblog text based on authorial style,” *Information Systems and e-Business Management*, vol. 15, no. 1, pp. 117–138, Feb. 2017, doi: 10.1007/s10257-016-0312-0.

[22] M. Altamimi and W. J. Teahan, “Gender and Authorship Categorisation of Arabic Text from Twitter Using PPM,” *International Journal of Computer Science and Information Technology*, vol. 9, no. 2, pp. 131–140, Apr. 2017, doi: 10.5121/ijcsit.2017.9212.

[23] M. Martinc, I. Škrjanec, K. Zupan, and S. Pollak, “PAN 2017: Author Profiling-Gender and Language Variety Prediction Notebook for PAN at CLEF 2017.” [Online]. Available: http://pan.webis.de/

[24] R. K. Bayot and T. Gonçalves, “Age and gender classification of tweets using convolutional neural networks,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 337–348. doi: 10.1007/978-3-319-72926-8_28.

[25] B. Bsir and M. Zrigui, “Bidirectional LSTM for author gender identification,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 393–402. doi: 10.1007/978-3-319-98443-8_36.

[26] S. Park and J. Woo, “Gender classification using sentiment analysis and deep learning in a health web forum,” *Applied Sciences (Switzerland)*, vol. 9, no. 6, 2019, doi: 10.3390/app9061249.

[27] R. Felipe, S. Dias, and I. Paraboni, “Cross-domain Author Gender Classification in Brazilian Portuguese,” 2020. [Online]. https://aclanthology.org/2020.lrec-1.154

[28] S. ElSayed and M. Farouk, “Gender identification for Egyptian Arabic dialect in twitter using deep learning models,” *Egyptian Informatics Journal*, vol. 21, no. 3, pp. 159–167, Sep. 2020, doi: 10.1016/j.eij.2020.04.001.

[29] L. Hilte, R. Vandekerckhove, and W. Daelemans, “Linguistic Accommodation in Teenagers’ Social Media Writing: Convergence Patterns in Mixed-gender Conversations,” *J Quant Linguist*, vol. 29, no. 2, pp. 241–268, 2022, doi: 10.1080/09296174.2020.1807853.

[30] T. K. Koch, P. Romero, and C. Stachl, “Age and gender in language, emoji, and emoticon usage in instant messages,” *Comput Human Behav*, vol. 126, Jan. 2022, doi: 10.1016/j.chb.2021.106990.

[31] C. Ikae and J. Savoy, “Gender identification on Twitter,” *J Assoc Inf Sci Technol*, vol. 73, no. 1, pp. 58–69, Jan. 2022, doi: 10.1002/asi.24541.

[32] B. Onikoyi, N. Nnamoko, and I. Korkontzelos, “Gender prediction with descriptive textual data using a Machine Learning approach,” *Natural Language Processing Journal*, vol. 4, p. 100018, Sep. 2023, doi: 10.1016/j.nlp.2023.100018.

[33] Manmeet Kaur, “A Comprehensive Overview of Artificial Intelligence-Based Classification Techniques,” *International Journal of Science and Research Archive*, vol. 11, no. 2, pp. 125–129, Mar. 2024, doi: 10.30574/ijsra.2024.11.2.0387.

[34] Y. K. Dwivedi *et al.*, “Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life,” *Int J Inf Manage*, vol. 55, Dec. 2020, doi: 10.1016/j.ijinfomgt.2020.102211.

[35] E. H. Almansor and F. K. Hussain, “Survey on Intelligent Chatbots: State-of-the-Art and Future Research Directions,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2020, pp. 534–543. doi: 10.1007/978-3-030-22354-0_47.

[36] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker, “Gender differences in language use: An analysis of 14,000 text samples,” May 2008. doi: 10.1080/01638530802073712.

[37] G. O. Olaoye, A. Luz, and E. Frank, “Machine Learning-based gender prediction with descriptive textual data.” [Online]. Available: https://www.researchgate.net/publication/378392987

[38] Yaman, Dogucan, et al. "Age and gender classification from ear images." *2018 International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2018.doi: 10.1109/IWBF.2018.8401568

[39] H. A. Alabbasi, F. Moldoveanu, and A. Moldoveanu, “HUMAN GENDER CLASSIFICATION USING KINECT SENSOR: A REVIEW,” *BAU Journal - Science and Technology*, vol. 5, no. 1, Dec. 2023, doi: 10.54729/2959-331x.1105.

[40] P. Thonglim, S. Thongsuwan, and P. Buranasiri, “Gender classification using convolutional neural networks based on fingerprint analysis with in-line digital holography,” SPIE-Intl Soc Optical Eng, Jan. 2024, p. 100. doi: 10.1117/12.3010005.

[41] M. M. Nasef, A. M. Sauber, and M. M. Nabil, “Voice gender recognition under unconstrained environments using self-attention,” *Applied Acoustics*, vol. 175, Apr. 2021, doi: 10.1016/j.apacoust.2020.107823.

[42] A. A. Alnuaim *et al.*, “Speaker Gender Recognition Based on Deep Neural Networks and ResNet50,” *Wirel Commun Mob Comput*, vol. 2022, 2022, doi: 10.1155/2022/4444388.

[43] Huy, Dinh Tran Ngoc, et al. "Further researches and discussion on machine learning meanings-and methods of classifying and recognizing users gender on internet." *Advances in Mechanics* 9.3 (2021): 1190-1204.

[44] Ö. ÇELİK and A. F. ASLAN, “Gender Prediction from Social Media Comments with Artificial Intelligence,” *Sakarya University Journal of Science*, vol. 23, no. 6, pp. 1256–1264, Dec. 2019, doi: 10.16984/saufenbilder.559452.

[45] M. Arshad, B. Khan, K. Khan, A. M. Qamar, and R. U. Khan, “ABMRF: An Ensemble Model for Author Profiling Based on Stylistic Features Using Roman Urdu,” *Intelligent Automation & Soft Computing*, vol. 0, no. 0, pp. 1–10, 2024, doi: 10.32604/iasc.2024.045402.

[46] M. Azhar *et al.*, “Real-Time Dynamic and Multi-View Gait-Based Gender Classification Using Lower-Body Joints,” *Electronics (Switzerland)*, vol. 12, no. 1, Jan. 2023, doi: 10.3390/electronics12010118.

[47] Z. Jiang, “Face gender classification based on convolutional neural networks,” in *Proceedings - 2020 International Conference on Computer Information and Big Data*

Applications, CIBDA 2020, Institute of Electrical and Electronics Engineers Inc., Apr. 2020, pp. 120–123. doi: 10.1109/CIBDA50819.2020.00035.

[48] Dong, Yujie, and Damon L. Woodard. "Eyebrow shape-based features for biometric recognition and gender classification: A feasibility study." *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 2011. DOI: 10.1109/IJCB.2011.6117511

[49] S. Ghosh and S. K. Setua, “Human gender revelation based on facial features and shape and structure of finger nails,” *Measurement and Control (United Kingdom)*, vol. 53, no. 7–8, pp. 1416–1428, Aug. 2020, doi: 10.1177/0020294020941873.

[50] Ueki, Kazuya, et al. "A method of gender classification by integrating facial, hairstyle, and clothing images." *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*. Vol. 4. IEEE, 2004.doi: 10.1109/ICPR.2004.1333798

[51] S. Belli and M. Jimenez, “Gender identity and emotions in email spam,” *Papeles del CEIC*, vol. 2015, no. 2, Sep. 2015, doi: 10.1387/pceic.13194.

[52] S. Dargan, M. Kumar, A. Mittal, and K. Kumar, “Handwriting-based gender classification using machine learning techniques,” *Multimed Tools Appl*, 2023, doi: 10.1007/s11042-023-16354-1.

[53] Mukherjee, Arjun, and Bing Liu. "Improving gender classification of blog authors." *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*. 2010.

[54] Z. M. Nia *et al.*, “Twitter-based gender recognition using transformers,” *Mathematical Biosciences and Engineering*, vol. 20, no. 9, pp. 15957–15977, 2023, doi: 10.3934/mbe.2023711.

[55] Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. "Gender identity and lexical variation in social media." *Journal of Sociolinguistics* 18.2 (2014): 135-160. https://doi.org/10.1111/josl.12080

[56] Zhao, Jieyu, et al. "Gender bias in coreference resolution: Evaluation and debiasing methods." *arXiv preprint arXiv:1804.06876* (2018).

[57] Blodgett, Su Lin, et al. "Language (technology) is power: A critical survey of" bias" in nlp." *arXiv preprint arXiv:2005.14050* (2020). https://doi.org/10.48550/arXiv.2005.14050

[58] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, “Semantic Similarity Metrics for Evaluating Source Code Summarization,” in *IEEE International Conference on Program Comprehension*, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/nnnnnnn.nnnnnnn.

[59] A. C. Kozlowski, M. Taddy, and J. A. Evans, “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings,” *Am Sociol Rev*, vol. 84, no. 5, pp. 905–949, Oct. 2019, doi: 10.1177/0003122419877135.

[60] Burger, John D., et al. "Discriminating gender on Twitter." *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011.

[61] Bergsma, Shane, et al. "Broadly improving user classification via communication-based name and location clustering on twitter." *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2013.

[62] T. P. Nguyen and A. C. Le, “A hybrid approach to Vietnamese word segmentation,” in *2016 IEEE RIVF International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for the Future, RIVF 2016 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Dec. 2016, pp. 114–119. doi: 10.1109/RIVF.2016.7800279.

[63] B. Mahesh, “Machine Learning Algorithms-A Review,” *International Journal of Science and Research*, 2018, doi: 10.21275/ART20203995.

[64] *CHAVAN, NIKHIL, and DEEPAK SHARMA. "2018 Fourth International Conference on Computing, Communication Control and Automation (ICCUBEA)." (2018).doi: 10.1109/ICCUBEA.2018.8697697*

[65] G. O. Olaoye and A. Luz, “Machine Learning Algorithms for Gender Prediction.February,2024. https://www.researchgate.net/publication/378499061

[66] M. Khan *et al.*, “Performance Evaluation of Machine Learning Models to Predict Heart Attack,” *Machine Graphics and Vision*, vol. 32, no. 1, pp. 99–114, 2023, doi: 10.22630/MGV.2023.32.1.6.

[67] P. Kumar, S. Arpan, K. Kar, Y. Singh, M. H. Kolekar, and S. Tanwar, “Lecture Notes in Electrical Engineering 597 Proceedings of ICRIC 2019 Recent Innovations in Computing.” Available: http://www.springer.com/series/7818

[68] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” May 01, 2021, *Springer*. doi: 10.1007/s42979-021-00592-x.

[69] Mining, W. I. D. (2006). Introduction to data mining (pp. 2-12). New Jersey: Pearson Education, Inc. https://doi.org/10.1007/978-1-4302-3325-1_14

[70] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Nov. 2018. doi: 10.1088/1742-6596/1142/1/012012.

[71] D. Swain, P. Kumar, P. Tushar, and A. Editors, “Advances in Intelligent Systems and Computing 1311 Machine Learning and Information Processing Proceedings of ICMLIP 2020.” [Online]. Available: http://www.springer.com/series/11156

[72] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[73] V. K. Giri, N. K. Verma, R. K. Patel, and V. P. Singh Editors, “Computing Algorithms with Applications in Engineering Algorithms for Intelligent Systems Series Editors: Jagdish Chand Bansal · Kusum Deep · Atulya K. Nagar.” [Online]. Available: http://www.springer.com/series/16171

[74] M. F. Hassan and M. E. Manaa, “Big Data Processing with Hadoop and Data Mining,” in *HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/HORA55278.2022.9800085.

[75] Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review, 1*(3), 9.

[76] M. A. Hambali, T. O. Oladele, K. S. Adewole, A. K. Sangaiah, and W. Gao, “Feature selection and computational optimization in high-dimensional microarray cancer datasets via InfoGain-modified bat algorithm,” *Multimed Tools Appl*, vol. 81, no. 25, pp. 36505–36549, Oct. 2022, doi: 10.1007/S11042-022-13532-5.

[77] R. Alkhatib, W. Sahwan, A. Alkhatieb, and B. Schütt, “A Brief Review of Machine Learning Algorithms in Forest Fires Science,” Jul. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app13148275.

[78] V. Kumar, “Evaluation of computationally intelligent techniques for breast cancer diagnosis,” *Neural Comput Appl*, vol. 33, no. 8, pp. 3195–3208, Apr. 2021, doi: 10.1007/s00521-020-05204-y.

[79] J. Gupta, S. Pathak, and G. Kumar, “Deep Learning (CNN) and Transfer Learning: A Review,” in *Journal of Physics: Conference Series*, Institute of Physics, 2022. doi: 10.1088/1742-6596/2273/1/012029.

[80] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00444-8.

[81] Panigrahi, Abhishek, Yueru Chen, and C-C. Jay Kuo. "Analysis on gradient propagation in batch normalized residual networks." (2018). https://doi.org/10.48550/arXiv.1812.00342

[82] Lorraine, Jonathan, and David Duvenaud. "Stochastic hyperparameter optimization through hypernetworks." *arXiv preprint arXiv:1802.09419* (2018).

[83] A. Ella Hassanien Roheet Bhatnagar Ashraf Darwish Editors, “Advances in Intelligent Systems and Computing 1141 Advanced Machine Learning Technologies and Applications Proceedings of AMLTA 2020.” http://www.springer.com/series/11156

[84] M. D. P. P. Goonathilake and P. P. N. V. Kumara, “Stance-Based Fake News Identification on Social Media with Hybrid CNN and RNN-LSTM Models,” *International Journal on Advances in ICT for Emerging Regions (ICTer)*, vol. 16, no. 3, pp. 1–12, Dec. 2023, doi: 10.4038/icter.v16i3.7234.

[85] Sakib, Shadman, et al. "An overview of convolutional neural network: Its architecture and applications." (2019). doi: 10.20944/preprints201811.0546.v4.

[86] M. Zulqarnain, R. Ghazali, Y. M. M. Hassim, and M. Rehan, “A comparative review on deep learning models for text classification,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 325–335, 2020, doi: 10.11591/ijeecs.v19.i1.pp325-335.

[87] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit Lett*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[88] S. Jugran, A. Kumar, B. S. Tyagi, and V. Anand, “Extractive Automatic Text Summarization using SpaCy in Python NLP,” in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 582–585. doi: 10.1109/ICACITE51222.2021.9404712.

[89] Sharma, Abhilasha, Raghav Aggarwal, and Raghav Alawadhi. "A Comparative Study of Text Summarization using Gensim NLTK Spacy and Sumy Libraries." *Journal of Xi’an Shiyou University, Natural Science Edition* 19 (2023).

[90] J. Han, M. Kamber, and J. Pei, “Getting to Know Your Data,” in *Data Mining*, Elsevier, 2012, pp. 39–82. doi: 10.1016/b978-0-12-381479-1.00002-2.

[91] Strang, Gilbert. *Introduction to linear algebra*. Wellesley-Cambridge Press, 2022.

[92] Axler, Sheldon. *Linear algebra done right*. Springer Nature, 2024.

[93] Grandini, Margherita, Enrico Bagli, and Giorgio Visani. "Metrics for multi-class classification: an overview." *arXiv preprint arXiv:2008.05756* (2020).

[94] Author, “Classification Performance Evaluation.”

[95] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep Learning-Based Text Classification,” Jun. 01, 2021, *Association for Computing Machinery*. doi: 10.1145/3439726.

[96] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model selection: A review with examples from ecology,” *Ecol Monogr*, vol. 93, no. 1, Feb. 2023, doi: 10.1002/ecm.1557.

[97] D. K. Sharma, M. Chatterjee, G. Kaur, and S. Vavilala, “Deep learning applications for disease diagnosis,” *Deep Learning for Medical Applications with Unique Data*, pp. 31–51, Jan. 2022, doi: 10.1016/B978-0-12-824145-5.00005-8.

[98] Mining, What Is Data. *Introduction to data mining*. New Jersey: Pearson Education, Inc, 2006.

[99] E. Sayilgan, Y. K. Yüce, and Y. İŞler, “Evaluation of mother wavelets on steady-state visually-evoked potentials for triple-command brain-computer interfaces,” *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 25, no. 9, pp. 2263–2279, 2021, doi: 10.3906/elk-2010-26.

[100] Y. Ho and S. Wookey, “The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling,” *IEEE Access*, vol. 8, pp. 4806–4813, 2020, doi: 10.1109/ACCESS.2019.2962617.

الخلاصة

إن الكم الهائل من البيانات النصية المتاحة في جميع أنحاء العالم، بما في ذلك المقالات ومحتوى وسائل التواصل الاجتماعي، قد أعطى أهمية لمنصات الوسائط مثل تويتر لاستخدام هذه البيانات ضمن تصنيف الجنس. هذا موضوع مثير للاهتمام للعديد من التطبيقات العملية مثل التسويق وأنظمة التوصية والجرائم الإلكترونية. يشير تصنيف الجنس في النص إلى عملية تصنيف الأفراد إلى أحد الجنسين، ذكر أو أنثى، بناءً على الخصائص اللغوية الملحوظة عادةً.

اكتسبت معالجة اللغة الطبيعية (NLP) شعبية في مجال التعلم الآلي. تطبق تقنيات معالجة اللغة الطبيعية (NLP) تصنيف الجنس تلقائيًا باستخدام السمات اللغوية والأسلوبية. يؤدي هذا إلى مشاركة أكبر ورضا، وتحسين دعم العملاء، وتقديم محتوى مخصص. الطبيعة الديناميكية والمفردات الضخمة للغة تجعل من الصعب تحديد جنس المؤلف بناءً على الأسلوب اللغوي، وقد كان هذا تحديًا للأطروحة، في حين أن طريقة استخراج الميزات المقترحة لها أهمية كبيرة في التغلب على هذه المشكلة وخلق تمييز دقيق بين الذكور والإناث.

الهدف من هذه الأطروحة هو تحسين دقة تصنيف الجنس بناءً على أسلوبه اللغوي في مجموعة بيانات المجال العام ومجموعة بيانات المجال المحدد. ولتحقيق هذا الهدف، تم استخراج الفروق النصية بين الجنسين باستخدام تشابه النص لتحسين تصنيف الجنس وتم تطبيق ثلاثة نماذج. تم تطبيق النموذج الأول من خلال ثلاث مصنفات للتعلم الآلي وهي الغابة العشوائية (RF) والانحدار اللوجستي (LR) وآلة المتجهات الداعمة (SVM) للحصول على تسميات الجنس واحتمالات التنبؤ الخاصة بها. وتم تطبيق النموذج الثاني من خلال الهياكل الناجحة المستخدمة مسبقًا لنماذج CNN للحصول على تسميات الجنس واحتمالاتها. أخيرًا، تم استخدام آلية التركيبة الخطية من خلال الجمع بين الأوزان الإضافية ونتائج احتمالية التسمية للنموذجين السابقين لحساب احتمالية التنبؤ النهائية.

تم الحصول على أعلى نتائج دقة لمجموعتي البيانات (Twitter وTripAdvisor). حقق نموذج التعلم الآلي 87.8٪ على Twitter، بينما حقق 75.1٪ على TripAdvisor. حصل نموذج التعلم العميق على 89.1٪ على Twitter و76.3٪ على TripAdvisor. أخيرًا، حقق نموذج التركيبة الخطية (89.6٪، 77٪) على Twitter وTripAdvisor على التوالي.

كان استخدام تقنية استخراج الميزات المقترحة أمرًا بالغ الأهمية في تحقيق نتائج متفوقة مقارنة بالأبحاث السابقة. بالإضافة إلى ذلك، فإن حقيقة أن مفردات مجموعة البيانات العامة لنطاق Twitter أكثر تنوعًا ساعدتنا على التفوق على TripAdvisor، وهي مجموعة بيانات نطاق محدد ذات دقة أقل بسبب لغتها المرتبطة فقط بالفنادق والمطاعم. من أجل تحقيق أقصى قدر ممكن من الدقة، كان استخدام استراتيجية التركيبة الخطية باستخدام التعلم العميق والتعلم الآلي مهمًا للغاية.

جامعة كربلاء
كلية علوم الحاسوب وتكنولوجيا المعلومات
قسم علوم الحاسوب

تصنيف الجنس بناءً على تحليل الأسلوب اللغوي باستخدام تقنيات التعلم الآلي والتعلم العميق

رسالة ماجستير
مقدمة الى مجلس كلية علوم الحاسوب وتكنولوجيا المعلومات / جامعة كربلاء وهي جزء من متطلبات نيل درجة الماجستير في علوم الحاسوب

كتبت بواسطة

حنين تاميم عبدعلي هاشم

بإشـراف

أ.م.د. ضمياء عباس حبيب

1446هـ **2024 م**