TECHNICAL UNIVERSITY OF SOFIA
ENGLISH LANGUAGE FACULTY OF ENGINEERING

DIPLOMA THESIS

EMOTION RECOGNITION MACHINE EARNING APPROACHES FOR DEPRESSION DETECTION

Submitted by:

Adrián Reviriego Reinaldo, 223223038

Supervisor:

MSc. Eng. Ralitza Raynova, PhD

Sofia, March 2024

Contents

Contents

Chapter 1

Introduction

1.1 Background

Depression has emerged as a critical issue in recent years, affecting millions of people worldwide. The global pandemic has exacerbated this situation, with increased isolation and stress contributing to a rise in depressive symptoms. The role of technology, particularly mobile phones and social media, has been double-edged, offering both support networks and contributing to the problem through increased exposure to negative content and social comparison. This section delves into the nature of depression and the significance of detecting it early.

1.1.1 Overview of Depression

Depression, historically referred to as melancholia, has been recognized as a significant mental health disorder for centuries. Ancient civilizations, such as the Greeks, acknowledged it as a medical condition, and treatments have evolved considerably over time. In the 20th century, significant advancements were made with the development of antidepressant medications and psychotherapy approaches. Despite these advancements, depression remains one of the leading causes of disability worldwide, affecting over 264 million people.

The COVID-19 pandemic and the widespread use of mobile phones and social media have further aggravated mental health challenges, especially among younger age groups. The increased screen time and exposure to online content can lead to negative social comparisons, cyberbullying, and reduced face-to-face interactions, all of which contribute to heightened levels of stress and depression. Understanding these modern contributors is essential in developing effective strategies to combat depression in today's digital age.

1.1.2 Importance of Early Detection

Early detection of depression is crucial for effective intervention and treatment. Recognizing the signs and symptoms of depression at an early stage allows for timely support and resources, which can significantly mitigate the impact on an individual's mental health. Early detection can lead to better treatment outcomes and can prevent the progression of depression to more severe stages. Tools and methods for early detection, such as screening questionnaires, digital monitoring, and AI-driven analysis of online behavior, are essential in identifying at-risk individuals before their condition worsens.

Timely intervention also reduces the overall burden on healthcare systems by addressing the issue before it escalates into more complex health problems. Schools, workplaces, and community settings are critical environments for implementing early detection programs. By fostering awareness and reducing the stigma associated with seeking help, early detection initiatives can encourage individuals to seek treatment sooner.

1.2 Significance

The significance of this research lies in its potential to enhance mental health screening and intervention strategies. By utilizing data from social media and other digital platforms, we develop more accurate and timely methods for identifying individuals at risk of depression. This section explores the broader implications of our findings and how they can contribute to better mental health outcomes.

1.2.1 Impact on Mental Health Screening

The analysis of social media platforms and digital communication channels plays a vital role in mental health screening and assessment. By leveraging data from social platforms, researchers and healthcare professionals can gain insights into prevailing sentiments, trends, and risk factors associated with depression. These insights enable targeted interventions and resource allocation to support individuals experiencing mental health challenges.

Innovative screening tools that analyze language patterns, social interactions, and activity levels on social media can identify individuals who may not seek help through traditional channels. These tools can flag potential signs of depression, such as changes in tone, increased use of negative language, or decreased social activity, prompting early intervention.

1.2.2 Potential for Early Intervention

Early intervention strategies hold promise in addressing the global burden of depression. By leveraging digital technologies and data-driven approaches, early intervention programs can identify individuals at risk and provide timely support and resources. Targeted interventions, such as online counseling services and digital mental health platforms, offer accessible and scalable solutions to promote mental well-being and prevent the onset of depressive symptoms.

Digital platforms can offer personalized interventions, such as cognitive-behavioral therapy modules, mindfulness exercises, and peer support networks, which can be accessed at the user's convenience. These interventions can be tailored to the individual's specific needs, making mental health care more responsive and effective. Additionally, integrating these digital tools with traditional healthcare services can create a comprehensive support system for individuals at risk of or suffering from depression.

Chapter 2

Purpose and Tasks

2.1 Introduction

In recent years, the intersection of machine learning (ML) and mental health research has emerged as a promising avenue for detecting and understanding depression. Depression, a prevalent and debilitating mental health disorder, affects millions of individuals worldwide, with significant societal and economic burdens [GFS+21]. Traditional methods of diagnosing depression often rely on self-reporting or clinician assessments, which can be subjective, time-consuming, and costly. In contrast, the application of machine learning techniques to detect depression offers the potential for objective, scalable, and efficient screening and monitoring tools.

This literature review aims to provide an overview of the advancements made in utilizing machine learning for the detection of depression. It will delve into the methodologies, datasets, and performance metrics employed in existing studies. Additionally, it will highlight the contributions of recent research and identify gaps and challenges in the field. Furthermore, this review will outline the proposed research direction and the intended contributions to the broader project of developing effective depression detection tools.

Through this literature review, we aim to provide insights into the state-of-the-art techniques, challenges, and future directions in the application of machine learning for depression detection. By addressing these aspects, we aim to contribute to the advancement of accurate, efficient, and accessible tools for detecting depression, ultimately improving mental health outcomes for individuals globally.

2.2 Depression Detection Methods

Depression is a complex mental health disorder characterized by persistent feelings of sadness, loss of interest or pleasure, changes in appetite or sleep patterns, fatigue, and difficulty concentrating. Diagnosing depression traditionally relies on clinical interviews, self-reported questionnaires, and observations by healthcare professionals [CSSWL21]. However, these methods have inherent limitations, including subjectivity, reliance on patient honesty, and variability in clinician expertise.

Clinical Interviews: Clinical interviews conducted by trained professionals are a common method for diagnosing depression. These interviews typically involve structured or semi-structured questions aimed at assessing symptoms and their severity. While clinical interviews allow for a comprehensive evaluation of an individual's mental state, they are time-consuming, costly, and reliant on the skill and experience of the interviewer.

Self-Reported Questionnaires: Self-reported questionnaires, such as the Beck Depression Inventory (BDI) [Upt13] and the Patient Health Questionnaire (PHQ-9) [Che22], are widely used screening tools for depression. These questionnaires consist of a series of statements or questions regarding depressive symptoms, and individuals rate their experiences on a scale. While self-reported questionnaires are convenient and easy to administer, they are susceptible to response biases, including social desirability bias and recall bias.

Observations by Healthcare Professionals: Healthcare professionals, including psychiatrists, psychologists, and primary care physicians, often rely on their clinical judgment and observations to diagnose depression. These observations may include changes in behavior, affect, and physical symptoms reported by the patient or observed during clinical encounters. However, reliance on observations alone can lead to misdiagnosis or underdiagnosis of depression, particularly in cases where symptoms are subtle or masked.

Despite the widespread use of these methods, challenges such as subjectivity, variability, and reliance on patient cooperation remain significant barriers to accurate depression detection. Additionally, the stigma associated with mental health disorders may deter individuals from seeking help or disclosing their symptoms, further complicating the diagnostic process.

Given these challenges, there is growing interest in leveraging machine learning techniques to improve the accuracy, efficiency, and accessibility of depression detection. Machine learning algorithms have the potential to analyze large volumes of data, identify patterns, and make predictions without human intervention. By utilizing diverse data sources, including electronic

health records, social media posts, smartphone sensor data, and neuroimaging scans, machine learning models can capture nuanced information and enhance our understanding of depression.

In the subsequent sections of this literature review, we will explore the advancements in machine learning for depression detection, including the methodologies, datasets, performance metrics, challenges, and proposed research directions. Through this exploration, we aim to elucidate the current landscape of machine learning applications in depression detection and identify opportunities for future research and development.

2.3 Advancements in Machine Learning for Depression Detection

Machine learning (ML) techniques have revolutionized the field of mental health research, particularly in the detection and diagnosis of depression. This section delves into recent advancements in ML applications for depression detection, drawing insights from a variety of studies and methodologies.

Deep Learning in Psychiatry: Squires et al. provide an extensive overview of AI and ML's role in reshaping psychiatry, with a specific focus on depression detection, diagnosis, and treatment. They note a significant shift towards advanced deep learning architectures and transformer-based embeddings for depression detection, emphasizing the need for more sophisticated techniques to improve treatment response prediction [STEea23].

Text Classification and Treatment Response Prediction: Varshney et al. offer a comprehensive exploration of AI and ML applications in detecting, diagnosing, and treating depression. They discuss advancements in text classification techniques and the prevalence of Support Vector Machines in treatment response prediction. Challenges such as small sample sizes and model validation issues are addressed, suggesting opportunities for future research [VGG22].

Speech Processing for Depression Prediction: Gaikwad et al. focus on predicting mental health-related problems, including depression, through speech processing. They highlight physiological changes in speech related to depression and propose leveraging signal processing techniques and statistical modeling for capturing natural changes in speech. Despite advancements, challenges such as limited dataset availability and model generalization persist [GV24].

Machine Learning on Socio-Demographic and Psychological Data: Mannepalli et al. analyze depression using socio-demographic and psychological data through ML techniques. Their study employs data preprocessing, feature selection, and classification techniques on a depression detection dataset. The research underscores the complexity of depression and the potential of ML for improving treatment outcomes [MKJ+24].

Facebook Data Analysis for Depression Detection: Islam et al. focus on detecting depression from Facebook posts using a multidimensional approach. They explore emotional, temporal, and linguistic features extracted from Facebook data and apply supervised ML techniques for depression detection. The study demonstrates the effectiveness of machine learning in identifying depressive behavior on social media [IKA+18].

Machine Learning with Face-to-Face Data Collection: Biswas et al. propose a system for detecting depression using machine learning algorithms and two feature selection approaches. Their research emphasizes the impact of depression on people's lives, particularly focusing on the COVID-19 situation, and advocates for real-life data collection methods [BIS+23].

Ethical Considerations in AI Technologies for Mental Health: Benrimoh et al. discuss the importance of clinical validation and ethical considerations in the development and implementation of AI technologies in mental health care. They emphasize the need for rigorous testing and transparency in AI development to ensure safety and efficacy [Bea18].

Psychological Impact of COVID-19 and Social Media Analysis: Arora et al. examine the psychological impact of the COVID-19 pandemic, particularly focusing on depression, anxiety, and stress. They discuss the utilization of social networking sites for identifying depressive mood disorders and the challenges associated with analyzing large volumes of social media data [AMKB21].

Machine Learning and Web-Based Application for Depression Detection: Shete et al. present a comprehensive approach to depression detection through machine learning and a web-based application. Their system achieves high accuracy in depression detection and treatment recommendation, offering a user-friendly interface for mental health assessment and support [SSB24].

Impact of COVID-19 on Mental Health: Iliou et al. examine the impact of the COVID-19 pandemic on the mental health of children and adolescents, focusing on depression, anxiety, and stress. They utilize machine learning classification algorithms to detect and categorize individuals based on demographic information and mental health responses [IKNea19].

Behavioral Indicators of Depression on Social Media: Shekerbekova et al. identify behavioral indicators of depression among users of the social network Vkontakte. They employ machine learning algorithms to detect depressive or suicidal behavior in online user content, emphasizing the significance of automating mental health projects and developing new forms of diagnosis [SYT+21].

Predicting Depression from Various Data Sources: Kilaskar et al. propose a system to predict depression by analyzing various parameters, including social media stories, tweets, voice recordings, and browser searches. They highlight the importance of leveraging machine learning algorithms for early detection and intervention strategies [KSAea22].

Predicting Prenatal Depression with Machine Learning: Preis et al. utilize machine learning methods to predict the risk of prenatal depression using data from the PROMOTE psychosocial screening tool. They demonstrate the potential of using machine learning to improve the screening and prediction of prenatal depression, aiming for more personalized interventions for maternal and infant health [PDAea22].

Comprehensive Approach to Depression Detection: Motade et al. propose a comprehensive approach to detecting depression using data analysis and machine learning techniques, leveraging datasets from the PHQ-9 questionnaire and Twitter. They emphasize the importance of early detection of depression and highlight the potential of machine learning approaches in this domain [MHMP22].

Hybrid Machine Learning Models for Depression Detection on Twitter: Khan and Alqahtani propose and evaluate multiple hybrid machine-learning models for sentiment analysis to detect signs of depression by analyzing Twitter tweets. They introduce four hybrid machine-learning models, evaluate their performance, and suggest policy implications for leveraging social media platforms for depression detection [KA23].

Predicting Depression in Home-Based Older Adults: Lin et al. develop a two-step hybrid machine learning model to predict the onset of depression in home-based older adults using longitudinal data. They discuss the methodology, results, and implications of the study, emphasizing the importance of considering time-varying predictors in depression prediction models [LWF22].

These studies collectively highlight the diverse methodologies and approaches employed in leveraging ML for depression detection. From neuroimaging and speech processing to social media analysis and predictive modeling, ML techniques offer promising avenues for enhancing our understanding of depression and improving diagnostic accuracy. However, challenges such

as dataset availability, model generalization, and ethical considerations remain areas for future research and development in this field.

2.4 Challenges and Limitations

While machine learning (ML) techniques offer promising avenues for enhancing depression detection, several challenges and limitations hinder the field's progress. These challenges span across various aspects, from data availability to ethical considerations.

Limited Data Availability: One of the primary challenges in ML-based depression detection is the availability of comprehensive and diverse datasets. Many studies face limitations due to small sample sizes, restricted access to clinical data, or imbalanced datasets. Limited data can lead to model overfitting, hindering the generalizability of results and potentially biasing the algorithms towards specific demographics or characteristics.

Model Generalization: Ensuring the generalizability of ML models across different populations and settings remains a significant concern. Models trained on specific datasets or populations may not perform effectively when applied to diverse or unseen data. Addressing this challenge requires robust validation strategies, including testing models on external datasets and assessing their performance across various demographic groups.

Ethical Considerations: Ethical considerations play a crucial role in the development and deployment of ML algorithms for depression detection. Privacy concerns, data security, and the potential misuse of sensitive information are key ethical challenges. Furthermore, biases inherent in training data or algorithmic decision-making can exacerbate existing disparities in healthcare outcomes, raising concerns about fairness and equity in algorithmic approaches.

Interpretability and Transparency: The black-box nature of many ML algorithms poses challenges in interpreting model decisions and understanding the factors driving predictions. Lack of interpretability can hinder clinical adoption and trust in algorithmic systems, particularly in healthcare contexts where decision-making transparency is essential. Developing interpretable models and ensuring transparency in algorithmic processes are critical steps in addressing this challenge.

Clinical Integration: Integrating ML-based depression detection tools into clinical practice poses practical challenges. Clinicians may require additional training to understand and interpret algorithmic outputs effectively. Moreover, integrating algorithmic tools into existing clinical workflows without disrupting care delivery presents logistical challenges. Successful

integration necessitates close collaboration between data scientists, clinicians, and healthcare administrators.

Bias and Fairness: ML algorithms trained on biased data can perpetuate or exacerbate existing disparities in healthcare outcomes. Biases may arise from historical inequalities in healthcare access, data collection practices, or algorithmic decision-making processes. Mitigating bias and ensuring fairness in ML-based depression detection require proactive measures, including careful dataset curation, bias assessment, and algorithmic fairness techniques.

Overall, addressing these challenges and limitations is essential for realizing the full potential of ML in depression detection. Collaborative efforts from researchers, clinicians, policymakers, and industry stakeholders are needed to overcome these hurdles and develop ethical, effective, and scalable ML solutions for improving mental health outcomes.

2.5 Proposed Research Direction and Contributions

Building upon the existing literature and recognizing the challenges and opportunities in applying machine learning (ML) for depression detection, this study aims to contribute to the field by focusing on specific research directions and methodologies tailored to address the identified gaps. Leveraging the CEASE corpus, which comprises phrases extracted from suicide notes, this research seeks to develop robust ML models for detecting depression with a particular emphasis on suicidal ideation.

Utilization of Specialized Data: One key direction of this research involves leveraging specialized datasets such as the CEASE corpus, which offers unique insights into the language and characteristics of individuals experiencing severe mental distress. By utilizing this corpus, the study aims to capture nuanced linguistic patterns associated with depression and suicidal ideation, thereby enhancing the accuracy and specificity of the ML models.

Feature Engineering and Model Development: The research will focus on comprehensive feature engineering techniques tailored to capture subtle linguistic cues indicative of depression and suicidal ideation. Features may include syntactic, semantic, and sentiment-based representations of text data extracted from the CEASE corpus.

Ethical Considerations and Bias Mitigation: The study will emphasize ethical considerations throughout the research process, particularly concerning patient privacy, data security, and algorithmic fairness. Transparent reporting of methodology and results will be prioritized to ensure accountability and trustworthiness of the developed models.

Overall, this research endeavors to advance the state-of-the-art in ML-based depression detection by leveraging specialized data, employing advanced modeling techniques, and prioritizing interpretability, generalization, and ethical considerations. By contributing novel insights and methodologies tailored to the detection of depression and suicidal ideation, this study aims to make meaningful contributions to the field of mental health diagnostics and ultimately improve patient outcomes.

2.6 Conclusion

In conclusion, the application of machine learning (ML) techniques for depression detection represents a promising avenue for improving mental health diagnostics and patient care. This literature review has provided insights into recent advancements, challenges, and potential research directions in this rapidly evolving field.

The reviewed studies have demonstrated the diverse methodologies and approaches employed in leveraging ML for depression detection, ranging from brain imaging and social media analysis to speech processing and predictive modeling. Despite the progress made, several challenges and limitations persist, including limited data availability, model generalization issues, ethical considerations, and biases inherent in algorithmic decision-making.

Moving forward, there is a need for continued interdisciplinary collaboration between researchers, clinicians, policymakers, and industry stakeholders to address these challenges effectively. By leveraging specialized datasets such as the CEASE corpus and employing advanced modeling techniques tailored to capture nuanced linguistic patterns associated with depression and suicidal ideation, future research can contribute to the development of robust and clinically relevant ML models for depression detection.

Moreover, ensuring the interpretability, generalizability, and ethical integrity of ML-based approaches will be paramount for their successful integration into clinical practice. By prioritizing transparency, accountability, and patient privacy throughout the research process, researchers can foster trust and confidence in algorithmic systems among healthcare professionals and patients alike.

Ultimately, the goal of ML-based depression detection research is to translate scientific advancements into tangible improvements in patient outcomes. By developing accurate, interpretable, and ethically sound ML models, researchers can empower clinicians with valuable tools for early detection, intervention, and personalized treatment of depression, thereby enhancing mental health care delivery and contributing to the well-being of individuals and communities.

Chapter 3

Chosen tool, technologies and libraries

3.1 Datasets

The selection and utilization of appropriate datasets are fundamental to the effectiveness of our research in Natural Language Processing (NLP). This study leverages multiple datasets to address tasks such as depression detection, sentiment analysis, and emotion recognition. The datasets used are described below:

The CEASE dataset is a specialized corpus focused on the analysis of suicide notes, aimed at understanding the mental state of individuals at risk of suicide. This dataset addresses the pressing public health concern of rising suicide rates by focusing on the detection of depression, a major factor contributing to suicidal ideation.

The CEASE dataset includes a standard emotion-annotated corpus of suicide notes in English, which has been extended with an additional 2,539 sentences collected from 120 new notes. Each sentence in the dataset is annotated with appropriate depression labels and multi-label emotion classes, providing a comprehensive framework for emotion recognition and mental health analysis.

Moreover, the dataset has been further enriched using weak supervision techniques to annotate the corpus with sentiment labels, adding another layer of annotation that supports various NLP tasks. This detailed annotation allows for in-depth analysis of the emotional and psychological states conveyed in the suicide notes, making the CEASE dataset a valuable resource for research focused on mental health and emotional well-being.

In addition to the CEASE dataset, two sentiment analysis datasets comprising thousands of tweets collected from the web were used. These datasets are crucial for understanding the

emotional tone and sentiment expressed in social media content, which is pivotal for tasks such as sentiment analysis and emotion detection.

Each tweet in these datasets is labeled with emotion tags, allowing for detailed sentiment classification and analysis. The datasets cover a wide range of emotions, providing a rich resource for evaluating sentiment analysis models. By leveraging these datasets, the research aims to build robust models capable of accurately classifying the emotional content of tweets, thereby contributing to a deeper understanding of public sentiment and emotional trends in social media.

3.1.1 Dataset Gathering

The datasets used in this research were obtained from various sources, each with specific procedures for access and compliance. Below, we detail the process for gathering and collecting each dataset.

CEASE Dataset: The CEASE dataset, which focuses on the analysis of suicide notes, was obtained directly from the authors of the study in which the dataset was originally created. To gain access, a formal request was made, followed by the completion of a form to ensure the dataset would be used for legitimate research purposes [GEB20].

Upon approval, the dataset was provided under the condition that the data compliance agreement would be strictly followed. This agreement outlines the ethical use of the data, including privacy considerations and restrictions on data sharing.

The dataset itself is organized into a set of text files, each corresponding to a specific emotion. For instance, a file named `blame_alltext.txt` contains phrases associated with the emotion of blame. This structured format facilitates the analysis of emotional expressions in the context of suicide notes.

Kaggle Dataset: Emotion Detection from Text: The second dataset used in this research was sourced from Kaggle, a well-known platform for data science competitions and datasets. This dataset, titled *Emotion Detection from Text*, is available for public use under specific licensing agreements provided by Kaggle.

The dataset is stored as a CSV file with three columns: `tweet_id`, `sentiment`, and `content`. Each row in the CSV file represents a tweet, with the `sentiment` column indicating the emotion expressed in the tweet and the `content` column containing the text of the tweet.

This dataset is instrumental for tasks involving sentiment analysis and emotion detection in social media content.

GitHub Dataset: The third dataset was obtained from a repository on GitHub, a platform widely used for hosting and sharing code and datasets. Similar to the Kaggle dataset, this dataset is also formatted as a CSV file and contains columns for `Emotion` and `Text`.

An additional version of this dataset includes a `Clean_text` column, which provides preprocessed text data. The columns in this version are `Emotion`, `Text`, and `Clean_text`. However, to maintain control over the preprocessing steps and ensure consistency with the methods employed in this research, the decision was made to preprocess the text data independently.

The acquisition of this dataset involved downloading it directly from the GitHub repository, which is publicly accessible. The use of this dataset complies with the licensing terms specified by the repository's authors.

3.2 Natural Language Processing (NLP) Tools

Natural Language Processing (NLP) is a crucial component of this research, enabling the analysis and understanding of human language data. In this section, we discuss the tools and techniques employed to process and analyze textual data. NLP is vital for our project as it underpins our ability to extract meaningful insights and detect depression indicators from the text.

3.2.1 Introduction to NLP

Natural Language Processing (NLP) is a field at the intersection of computer science, artificial intelligence, and linguistics, focused on the interaction between computers and human (natural) languages. The goal of NLP is to enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful [KKK+23].

NLP encompasses a wide range of tasks, including but not limited to:

- **Text Classification:** Assigning predefined categories to text.
- **Sentiment Analysis:** Identifying and categorizing opinions expressed in text.
- **Named Entity Recognition (NER):** Detecting and classifying proper nouns in text.

- **Machine Translation:** Automatically translating text from one language to another.
- **Speech Recognition:** Converting spoken language into text.
- **Tokenization:** Breaking text into smaller units, such as words or phrases.
- **Parsing:** Analyzing the grammatical structure of a sentence.

The applications of NLP are vast and varied, ranging from simple text processing and sentiment analysis to sophisticated conversational agents and machine translation systems. NLP techniques are instrumental in making sense of large volumes of unstructured text data, enabling advancements in fields such as customer service, healthcare, finance, and many others.

3.2.2 Chosen NLP Tools

In this research, various NLP tools and libraries were utilized to preprocess text data, perform sentiment analysis, and build machine learning models for natural language processing tasks. The following NLP tools were chosen for their effectiveness and suitability for the specific tasks at hand:

- **NLTK (Natural Language Toolkit):** NLTK is a comprehensive library for NLP tasks in Python, providing functionalities such as tokenization, lemmatization, stemming, and sentiment analysis. It is widely used for text preprocessing and feature extraction [LB02].
- **Contractions:** Contractions is a Python library for expanding contractions in text data. It replaces contractions (e.g., "can't" to "cannot") to improve text normalization and analysis.
- **Num2Words:** Num2Words is a Python library for converting numbers into words. It is useful for preprocessing text data containing numerical information, such as converting numeric values into textual representations.
- **VADER (Valence Aware Dictionary and Sentiment Reasoner):** VADER is a lexicon and rule-based sentiment analysis tool specifically designed for social media texts. It provides a pre-trained model for analyzing sentiment polarity in text data [SSRM23].

These chosen NLP tools and libraries were instrumental in preprocessing text data, extracting features, performing sentiment analysis, and building machine learning models for various NLP tasks in this research.

3.3 Rationale for Choosing Python

Python has become the leading programming language for machine learning and natural language processing (NLP) due to its simplicity, versatility, and the extensive ecosystem of libraries and tools. Its ease of use, rich library support, and strong community make it an ideal choice for researchers and developers in these fields. In this research, we leverage Python's capabilities to perform various NLP and machine learning tasks. The following sections will discuss Python's relevance to these domains and provide an overview of the specific libraries and frameworks employed in this project [Mat19].

3.3.1 Python's Relevance to Machine Learning and NLP

Python's prominence in the field of machine learning and natural language processing (NLP) is well-established, owing to its extensive capabilities, ease of use, and robust library ecosystem. Several key factors highlight the relevance of Python in machine learning:

- **Comprehensive Libraries and Frameworks:** Python offers a rich array of libraries and frameworks that are specifically designed for machine learning and NLP tasks. Libraries such as Scikit-learn, TensorFlow, and PyTorch provide powerful tools for model building, training, and evaluation. Additionally, NLP-specific libraries like NLTK and spaCy streamline the processing and analysis of textual data.
- **Ease of Implementation:** Python's intuitive syntax and high-level programming constructs make it accessible and easy to implement complex machine learning algorithms. This simplicity accelerates the development cycle, allowing researchers and developers to quickly prototype and iterate on their models.
- **Extensive Community and Support:** Python benefits from a large and active community of developers, researchers, and machine learning practitioners. This community-driven ecosystem ensures continuous improvement of libraries, comprehensive documentation, and a wealth of tutorials and resources for troubleshooting and learning.
- **Integration Capabilities:** Python integrates seamlessly with other technologies and platforms, facilitating the incorporation of machine learning models into larger systems and workflows. This interoperability is crucial for deploying machine learning solutions in real-world applications.

- **Versatile Applications:** Python's versatility extends to various domains beyond machine learning and NLP, including data analysis, web development, scientific computing, and automation. This broad applicability makes Python a valuable tool for interdisciplinary research and development.

In summary, Python's comprehensive libraries and frameworks, ease of implementation, strong community support, integration capabilities, and versatile applications underscore its relevance and effectiveness in advancing machine learning and NLP research.

3.3.2 Python Libraries and Frameworks

In this research, several Python libraries and frameworks were utilized to perform various tasks related to natural language processing (NLP) and machine learning. Each library was selected for its specific capabilities and contributions to the overall project. The key libraries and frameworks used include:

- **NLTK (Natural Language Toolkit):** A comprehensive library for NLP tasks such as tokenization, stemming, lemmatization, and sentiment analysis. NLTK is fundamental for initial text processing and exploratory analysis.
- **Num2Words:** This library is used to convert numerical values into their textual representations, enhancing the normalization process of textual data.
- **Contractions:** Contractions is used to expand contracted forms in the text, which aids in the standardization and preprocessing of the data.
- **WordCloud:** A visualization tool that creates word clouds from text data, helping to visualize word frequencies and patterns, which is useful for gaining insights into the most common terms in the dataset.
- **Seaborn and Matplotlib:** These are powerful visualization libraries used for creating various plots and charts to illustrate data distributions, relationships, and trends. They are essential for data exploration and the presentation of research findings [Bis19].
- **Pandas and NumPy:** These libraries are fundamental for data manipulation and numerical computations. Pandas provides data structures like DataFrames, while NumPy supports efficient numerical operations on large datasets [PJ22].

- **Scikit-learn:** Although primarily used for machine learning tasks, Scikit-learn also provides tools for data preprocessing, feature extraction, and evaluation. It includes implementations of various algorithms that are applied to text data [Pap20].
- **VaderSentiment:** Part of NLTK, the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool is specifically tuned for analyzing social media text, providing nuanced sentiment analysis capabilities.
- **WordNetLemmatizer and PorterStemmer:** These tools from NLTK are used for lemmatization and stemming, respectively, helping to reduce words to their base or root forms for more effective analysis.

These libraries and frameworks were chosen for their robustness, ease of use, and extensive functionalities, which significantly facilitated the development and implementation of the NLP and machine learning solutions in this research.

Chapter 4

Description of the methodology

4.1 Dataset Acquisition

The first essential step in developing our research is to load the datasets into the Python environment. This process involves reading the data from various sources and ensuring it is in a suitable format for analysis. Proper data acquisition is crucial for the integrity and success of subsequent data processing and analysis steps. In this section, we will detail how the datasets were uploaded and structured for use in our research.

4.1.1 Data Loading

To begin our analysis, we needed to upload three distinct datasets into the Python environment. Each dataset was handled using different methods tailored to their specific formats and storage locations.

The CEASE dataset, which consists of multiple text files each representing different emotions, was uploaded using the `os` library. This approach allowed us to read all files in the dataset directory and store the contents efficiently. The files were read into memory and their contents were stored in a pandas DataFrame for easy manipulation and analysis.

The other two datasets were sourced from Kaggle and GitHub, respectively. Both of these datasets were in CSV format and were uploaded using the `pandas` library. The `pandas` library provides robust functions for reading CSV files, allowing us to load the data quickly and store it in DataFrames. This method ensures that the data is structured and ready for preprocessing and analysis.

All three datasets, once loaded, were stored in pandas DataFrames. This storage format facilitates efficient data manipulation and supports the various preprocessing techniques and analyses required for our research.

4.1.2 Data Restructuring

After loading the datasets into the Python environment, the next step was to restructure the data to suit our analysis needs. This involved converting all datasets from their respective formats to a unified format with two columns: `Text` and `label`.

Using `pandas`, we transformed each dataset to have a consistent structure. The `Text` column contains the phrases or text entries from the datasets, while the `label` column contains the corresponding emotion labels.

Additionally, we simplified the labeling problem from a multilabel to a binary classification task. We separated the positive emotions from the negative ones, converting all positive emotions to the label `1` and all negative emotions to the label `-1`. This binary labeling approach allows us to classify emotions into positive and negative categories, which is more straightforward for the machine learning models we intend to use.

As a result, we obtained three datasets, each with the columns `Text` and `label`, where the `Text` column contains the phrases and the `label` column contains `-1` or `1` depending on whether they represent negative or positive emotions.

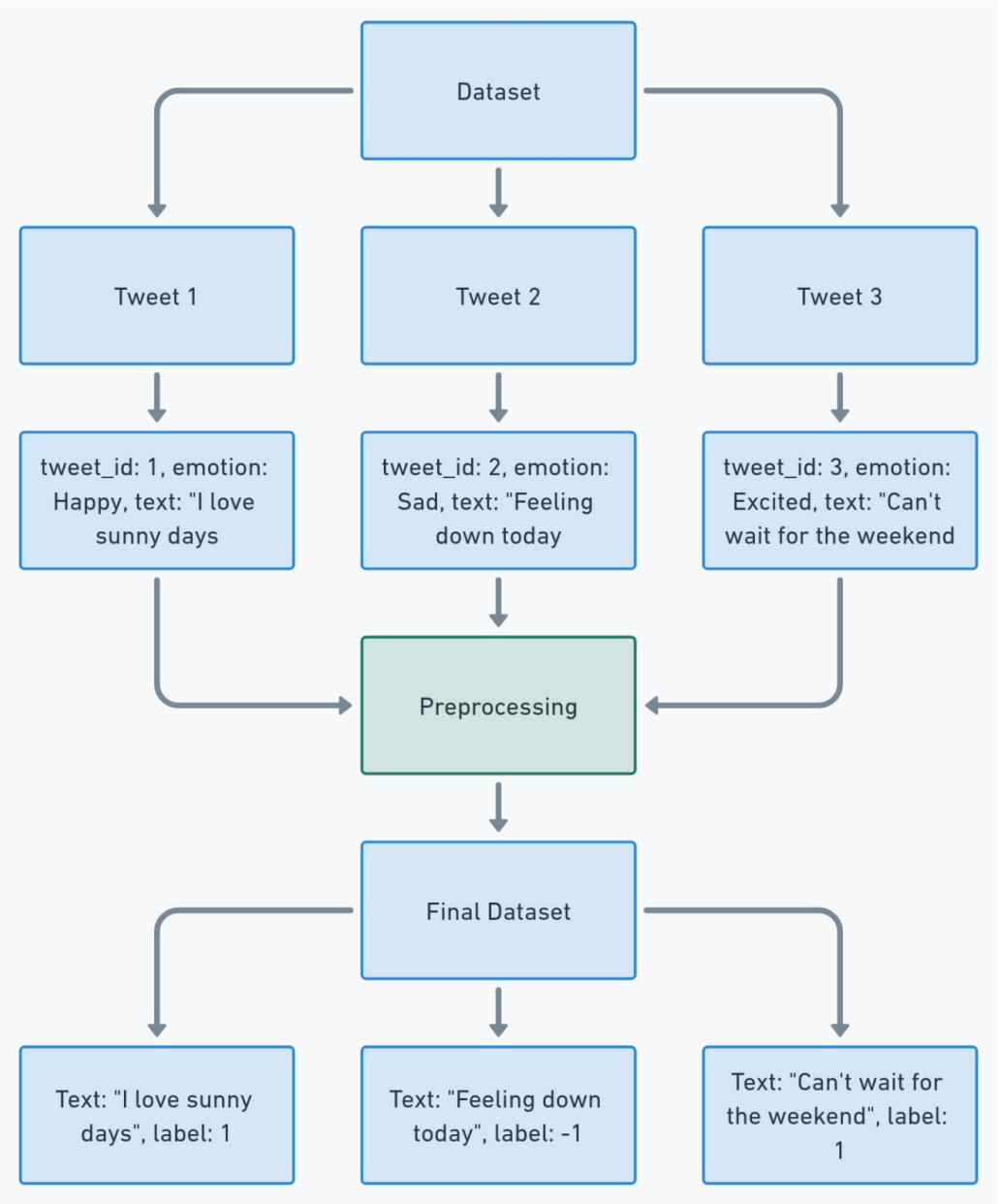

FIGURE 4.1: Dataset reestructuring and label preprocessing.

4.2 Data Preprocessing

Data preprocessing is a crucial step in the data analysis pipeline, as it ensures that the data is clean, consistent, and ready for modeling. This stage involves several tasks, including data cleaning and text processing, which help to enhance the quality and usability of the data. To understand the dataset better, we generated word clouds for each dataset. A word cloud is a visual representation of text data, where the size of each word indicates its frequency

or importance. This visualization helps in identifying the most common terms and gives an overview of the dataset's content[BFK+14].

4.2.1 Data Cleaning

Data cleaning involves removing or correcting corrupt or inaccurate records from the dataset. The following functions were employed to clean the data:

- **remove_non_ascii(text):** This function removes non-ASCII characters from a list of tokenized words. ASCII characters are standard characters in the English language, and removing non-ASCII characters helps in maintaining uniformity.
- **remove_punctuation(text):** This function removes punctuation from a list of tokenized words. Punctuation marks do not contribute to the meaning in a way that is beneficial for most text analysis tasks.
- **remove_url(text):** This function removes URLs from a list of tokenized words, as URLs typically do not add useful information for text analysis in the context of sentiment or emotion detection.
- **remove_consecutive_characters(text):** This function removes consecutive characters that appear three or more times in a word, which are often typing errors or stylized text that do not contribute meaningfully to analysis.
- **remove_usernames(text):** This function removes usernames appearing in a list of tokenized words, as usernames do not add value to the sentiment or emotion analysis.
- **remove_frequent_words(text):** This function removes frequently occurring words that do not contribute meaningful information to the analysis, ensuring that only relevant words are considered.
- **remove_stopwords(words):** This function removes stopwords, which are common words that do not significantly influence the model, from a list of tokenized words. Stopwords include words like "the", "is", and "and".

4.2.2 Text Processing

Text processing involves transforming the text into a format that can be effectively used for analysis. The following functions were employed for text processing:

- **to_lowercase(text):** This function converts all words to lowercase in a list of tokenized words. This standardization helps in treating words like "The" and "the" as the same word.
- **replace_numbers(text):** This function replaces all occurrences of numbers with their textual representation in a list of tokenized words. This helps in treating numeric data consistently.
- **lemmatize(text):** This function lemmatizes words in each text. Lemmatization involves obtaining the lemma of a word, which is its base or dictionary form. For example, words like "eating" and "eats" when lemmatized would yield "eat".
- **stemmatize(text):** This function stems words in each text, which involves shortening words to their root form. For example, words like "eating" and "eats" when stemmed would yield "eat".
- **normalize(text):** This function applies multiple steps to normalize the text for effective learning. It includes removing non-ASCII characters, usernames, URLs, converting to lowercase, removing punctuation, replacing numbers, removing consecutive characters, removing stopwords, and removing frequent words.

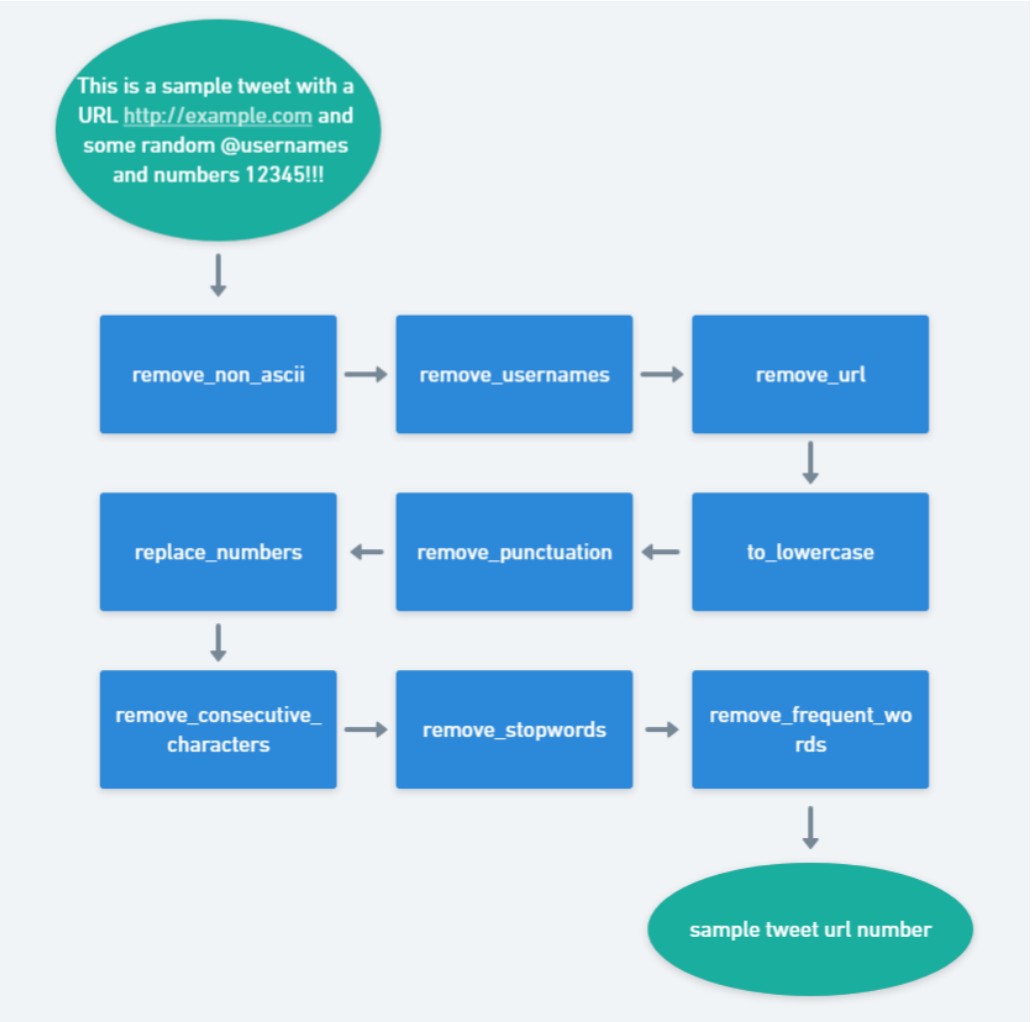

FIGURE 4.2: Flow of a tweet going through the preprocessing phase.

4.3 Model Creation

Model creation is a pivotal stage in the data analysis process, where we develop machine learning models to predict emotions or sentiments based on textual data. In this section, we discuss the selection of models, construction of pipelines, and hyperparameter tuning to optimize model performance.

4.3.1 Model Selection

The selection of appropriate machine learning models is crucial for achieving accurate and reliable predictions. We have chosen a variety of models tailored to the task of sentiment or

emotion analysis:

- **Naïve Bayes:** Naïve Bayes classifiers are simple yet effective probabilistic classifiers based on Bayes' theorem with the "naïve" assumption of independence among features. They are particularly suitable for text classification tasks and perform well with high-dimensional data [Web10].
- **Logistic Regression:** Logistic regression is a linear model used for binary classification tasks. Despite its simplicity, logistic regression can capture complex relationships between features and target variables, making it suitable for sentiment analysis tasks [Das14].
- **Neural Networks:** Neural networks, especially deep learning models, have shown remarkable success in various natural language processing tasks, including sentiment analysis. They can learn intricate patterns and relationships from textual data, leading to high predictive performance [Brä03].
- **Support Vector Machines (SVM):** SVM is a powerful supervised learning algorithm capable of performing classification tasks. SVM constructs hyperplanes in a high-dimensional space to separate data points into different classes, making it effective for sentiment analysis tasks [CS08].
- **Random Forest:** Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is robust against overfitting and works well with high-dimensional data [Bre01].
- **AdaBoost:** AdaBoost, short for Adaptive Boosting, is an ensemble technique that combines the outputs of several weak classifiers to produce a strong classifier. It adjusts the weights of incorrectly classified instances, emphasizing harder cases, which improves the model's performance [CHB17].
- **One-vs-One (OVO) and One-vs-All (OVA):** OVO and OVA are strategies for extending binary classification algorithms to multiclass classification problems. OVO constructs a classifier for every pair of classes, while OVA builds multiple binary classifiers, each distinguishing between one class and the rest. These strategies are useful for handling datasets with multiple emotion categories [LGT+22].

Each of these models has its unique advantages and may perform differently depending on the characteristics of the dataset and the nature of the problem. By experimenting with multiple models, we aim to identify the best-performing one for our specific sentiment or emotion analysis task.

4.3.2 Pipeline Construction

To understand the pipeline construction, it's essential to first grasp the concept of a vectorizer. In natural language processing (NLP), a vectorizer is a tool used to convert text data into numerical vectors that machine learning models can understand. This process is crucial because most machine learning algorithms require numerical input for training and prediction [DAU22].

There are several types of vectorizers, but two commonly used ones are CountVectorizer and TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer.

- **CountVectorizer**: This vectorizer converts a collection of text documents into a matrix of token counts. It counts the frequency of each word occurring in the document. Each document's content is represented as a vector, where each element of the vector corresponds to the count of a specific word in the document.
- **TF-IDF Vectorizer**: TF-IDF vectorizer, on the other hand, considers not only the frequency of a word in a document but also its importance in the entire corpus. It calculates a weight for each word based on its frequency in the document (Term Frequency, TF) and its rarity across all documents in the corpus (Inverse Document Frequency, IDF). Words that appear frequently in a specific document but rarely across other documents are considered more important.

Both vectorizers have their advantages and are suitable for different scenarios. CountVectorizer is simple and computationally efficient, making it a good choice for basic NLP tasks. TF-IDF vectorizer, on the other hand, takes into account the uniqueness of words across the entire corpus, which can be beneficial in capturing the importance of terms in a document.

Now, let's discuss the pipeline construction. In machine learning, a pipeline is a sequence of data processing components arranged sequentially, where the output of one component serves as the input to the next. In our context, a pipeline typically consists of two main components: a vectorizer and a model [BJV+12].

- **Vectorizer**: As mentioned earlier, the vectorizer prepares the text data by converting it into numerical vectors using techniques like CountVectorizer or TF-IDF vectorizer. It processes the raw text, tokenizes it, and performs other preprocessing steps such as removing stopwords, stemming, or lemmatization.
- **Model**: The model is the machine learning algorithm that learns patterns from the vectorized text data and makes predictions. It could be any of the models we discussed

earlier, such as Naïve Bayes, Logistic Regression, Neural Networks, SVM, Random Forest, AdaBoost, or one of their variants like OVO or OVA classifiers.

During the training phase, the pipeline takes the raw text data as input, applies the vectorizer to convert it into numerical vectors, and then trains the model on these vectors. In the prediction phase, the pipeline applies the same preprocessing steps to new incoming text data and feeds it into the trained model to make predictions.

By constructing pipelines in this manner, we can streamline the entire process of text data preprocessing and model training, making it easier to experiment with different models and preprocessing techniques while ensuring consistency and reproducibility.

4.3.3 Hyperparameter Tuning

Choosing the right hyperparameters for a machine learning model is crucial for achieving optimal performance. Hyperparameters are parameters that are not learned directly from the data but are set prior to the learning process. They control aspects such as the complexity of the model, regularization strength, and optimization strategy. Tuning these hyperparameters involves finding the combination that yields the best performance on a validation set.

Hyperparameters of Vectorizer

Before delving into the hyperparameters of each model, let's first discuss the hyperparameters of the vectorizer. In our case, we are using either CountVectorizer or TF-IDF vectorizer. The main hyperparameters to consider for these vectorizers are:

- **max_features**: This parameter specifies the maximum number of features (words) to extract from the text. It helps control the dimensionality of the feature space.
- **ngram_range**: This parameter determines the range of n-grams (sequences of adjacent words) to consider. For example, (1, 1) represents unigrams, (1, 2) represents unigrams and bigrams, and so on.

Now, let's proceed to discuss the hyperparameters specific to each model.

Hyperparameters of Naïve Bayes

For Naïve Bayes, the hyperparameters are:

- **alpha**: Laplace smoothing parameter.
- **fit_prior**: Whether to learn class prior probabilities from the data or to use a uniform prior.

Hyperparameters of Logistic Regression

For Logistic Regression, the hyperparameters are:

- **C**: Inverse of regularization strength.
- **solver**: Optimization algorithm.

Hyperparameters of Neuronal Network

For Neural Networks, the hyperparameters are:

- **hidden_layer_sizes**: Number of neurons in each hidden layer.
- **activation**: Activation function.
- **alpha**: L2 penalty (regularization term).

Hyperparameters of SVM

For SVM, the hyperparameters are:

- **C**: Regularization parameter.
- **kernel**: Kernel function.
- **gamma**: Kernel coefficient.

Random Forest:

For Random Forest, the hyperparameters are:

- **n_estimators**: Number of trees in the forest.
- **max_depth**: Maximum depth of the tree.

- **min_samples_split**: Minimum number of samples required to split an internal node.
- **min_samples_leaf**: Minimum number of samples required to be at a leaf node.

AdaBoost:

For AdaBoost, the hyperparameters are:

- **n_estimators**: Number of base learners.
- **learning_rate**: Learning rate shrinks the contribution of each classifier.

Hyperparameters of OVO and OVA

For OVO and OVA, the hyperparameters are similar to SVM:

- **C**: Regularization parameter.
- **kernel**: Kernel function.
- **gamma**: Kernel coefficient.

To determine the optimal values for each model, we employed GridSearchCV, a method that exhaustively searches through a specified parameter grid to find the best combination of hyperparameters. We used 5-fold cross-validation to evaluate the performance of each set of hyperparameters. Cross-validation is a robust technique for assessing the generalizability of a model. In 5-fold cross-validation, the dataset is divided into five equal parts. The model is trained on four parts and tested on the remaining part. This process is repeated five times, with each part used exactly once as the test set. The final performance metric is the average of the five test results, providing a more reliable estimate of the model's performance compared to a single train-test split.

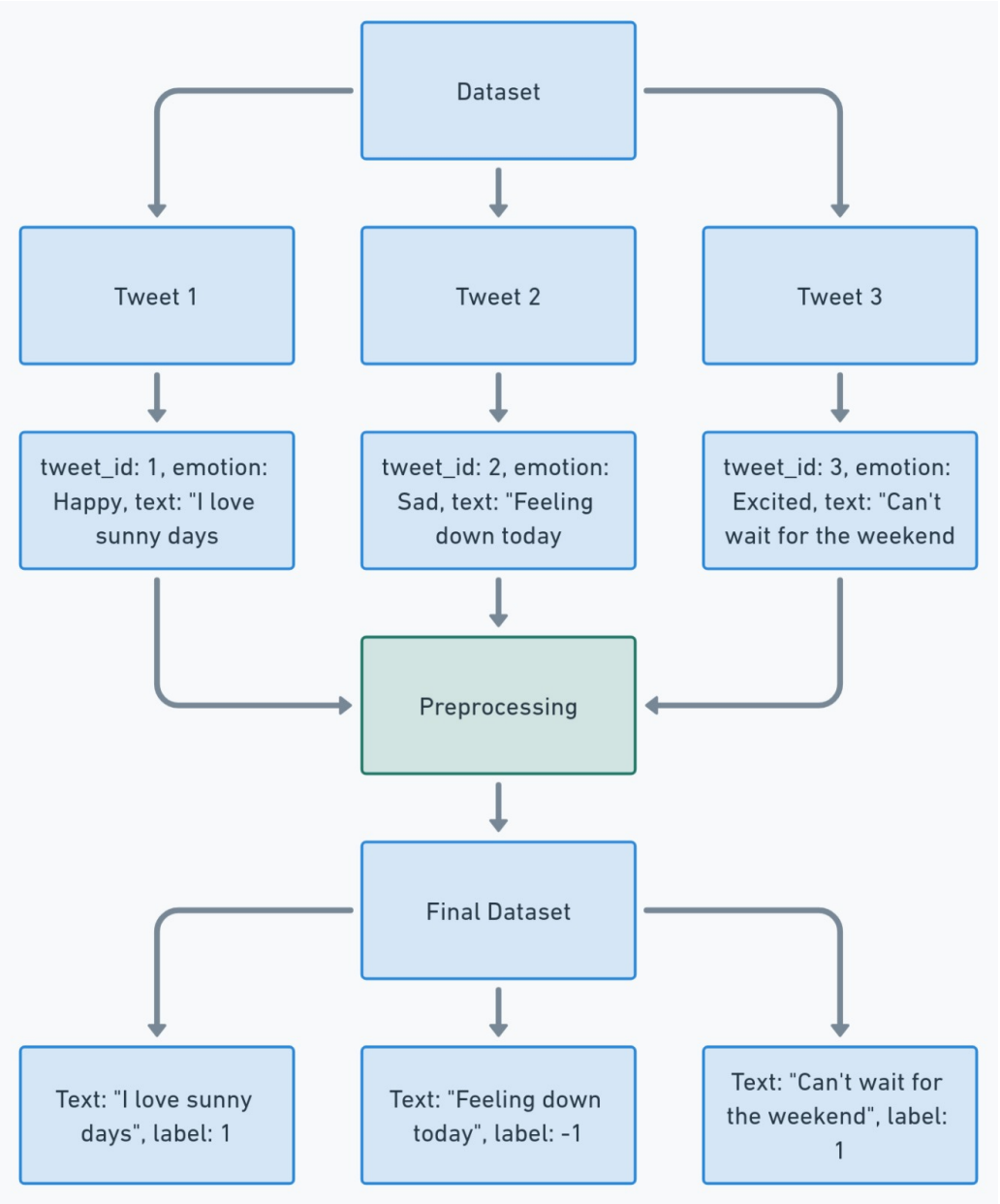

FIGURE 4.3: Overview of the project process.

4.4 Results Analysis

This section focuses on the metrics and visualizations used to evaluate the models' performance. The metrics provide quantitative measures of the models' effectiveness, while the visualizations offer intuitive insights into their predictive capabilities. Detailed results will be presented in subsequent sections.

4.4.1 Performance Metrics

Performance metrics are essential for evaluating the effectiveness of machine learning models. In this analysis, we use five key metrics: accuracy, precision, recall, F1-score, and ROC-AUC score [VC23].

Accuracy Accuracy measures the proportion of correctly predicted instances out of the total instances. It is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Accuracy provides a general idea of how well a model is performing. However, it may not be the best metric for imbalanced datasets, where one class is significantly more prevalent than the other.

Precision Precision, also known as positive predictive value, measures the proportion of positive identifications that were actually correct. It is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Precision is important in scenarios where the cost of false positives is high. It helps in understanding how many of the predicted positives are actual positives.

Recall Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives that are correctly identified by the model. It is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall is crucial for problems where identifying all positive instances is more important than minimizing false positives. This metric is particularly relevant in our context, as missing a positive instance could have significant implications.

F1-Score The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is useful for assessing the balance between precision and recall, especially in cases of imbalanced classes. It provides a more comprehensive measure than either precision or recall alone.

ROC-AUC Score The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) score measures the ability of the model to distinguish between classes. The ROC curve is a plot of the true positive rate against the false positive rate at various threshold settings. The AUC score represents the degree or measure of separability achieved by the model. It is defined as:

$$\text{ROC-AUC} = \int_0^1 \text{ROC}(t)\, dt$$

where ROC(t) is the true positive rate at threshold t. The ROC-AUC score ranges from 0 to 1, with 1 indicating perfect classification and 0.5 indicating no better than random chance. This metric is particularly useful for evaluating model performance across different threshold values and provides a comprehensive picture of the model's discriminative ability.

By considering these metrics, we can gain a thorough understanding of our model's performance across different aspects and ensure a robust evaluation.

4.4.2 Visualizations

Visualizations play a vital role in interpreting and comparing model performance. They help in understanding the distribution of predictions and identifying potential areas for improvement.

Confusion Matrices A confusion matrix is a table used to describe the performance of a classification model. It shows the actual versus predicted classifications, providing insight into the types of errors the model makes. The confusion matrix includes the following elements:

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positives (TP)	False Negatives (FN)
Actual Negative	False Positives (FP)	True Negatives (TN)

Table 4.1: Confusion Matrix

Analyzing the confusion matrix helps identify whether the model is more prone to false positives or false negatives, guiding further model tuning and improvement.

Graphs Comparing Model Performance In addition to confusion matrices, we use various graphs to compare the performance metrics of different models. These graphs include:

- **Accuracy:** A graph showing the overall correctness of each model.
- **Precision:** A graph illustrating the proportion of positive identifications that were actually correct for each model.

- **Recall:** A graph displaying the proportion of actual positives that were correctly identified for each model.
- **F1-Score:** A graph presenting the harmonic mean of precision and recall for each model.
- **ROC-AUC:** A graph comparing the area under the ROC curve for each model.

These visualizations provide a comprehensive comparison of all metrics across various models, helping us determine which model performs best overall. By analyzing these graphs, we can make informed decisions about model selection and optimization.

Overall, the combination of performance metrics and visualizations provides a thorough understanding of model performance, guiding further improvements and ensuring robust and reliable results.

Chapter 5

Results

5.1 Model Results

In this section, we present the culmination of our efforts in developing and evaluating machine learning models for the detection of depression. Building upon the foundation laid out in the preceding sections, where we detailed the methodologies and datasets utilized, we now delve into the outcomes of our computational analyses.

Our primary objective is to assess the performance of various machine learning algorithms in distinguishing between individuals with depression and those without. Through rigorous experimentation and meticulous evaluation, we aim to discern the strengths and weaknesses of each model, thereby shedding light on their utility in real-world applications.

The structure of this section is designed to provide a comprehensive overview of our findings. We begin by reporting the performance metrics of individual models, including accuracy, precision, recall, F1-score, and ROC-AUC score. Subsequently, we undertake a comparative analysis, elucidating the relative efficacy of different algorithms and identifying potential avenues for improvement.

Finally, beyond mere numerical outcomes, we endeavor to offer insights into the interpretability and generalizability of the models. By dissecting the underlying mechanisms driving model predictions and scrutinizing factors influencing performance, such as dataset characteristics and the prevalence of false positives and false negatives, we aim to provide a nuanced understanding of our results.

5.1.1 Naïve Bayes

The Naïve Bayes classifier, known for its simplicity and efficiency, presents a compelling approach to depression detection through machine learning. Leveraging probabilistic principles, Naïve Bayes models offer unique insights into the nuanced patterns within depression-related data.

Table 5.1: Performance Metrics on Different Datasets

Dataset	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
CEASE	76.5%	76.9%	97.2%	85.9%	58.1%
Tweets	60.8%	60.9%	96.8%	74.8%	51.8%
Other	49.6%	48.6%	97.2%	64.7%	51.7%

The table summarizes the performance metrics of the Random Forest classifier on three different datasets: CEASE, Tweets, and another dataset.

Overall, the performance of the classifier varies across different datasets. On the CEASE dataset, the classifier achieves the highest accuracy of 76.5% with relatively balanced precision and recall scores. However, the ROC-AUC score indicates relatively poor discrimination ability.

On the Tweets dataset, the classifier exhibits moderate accuracy (60.8%) and precision (60.9%), but high recall (96.8%), suggesting a high sensitivity to individuals with depression. The F1-score indicates a reasonable balance between precision and recall, although the ROC-AUC score is relatively low.

On another dataset, the classifier's performance is less satisfactory, with an accuracy of 49.6% and a high rate of false positives (48.6%). Despite this, the classifier maintains a high recall rate (97.2%), indicating its effectiveness in identifying individuals with depression. However, similar to the Tweets dataset, the ROC-AUC score suggests limited discrimination ability.

These findings underscore the importance of evaluating classifier performance across multiple metrics and datasets to gain a comprehensive understanding of its capabilities and limitations in real-world scenarios.

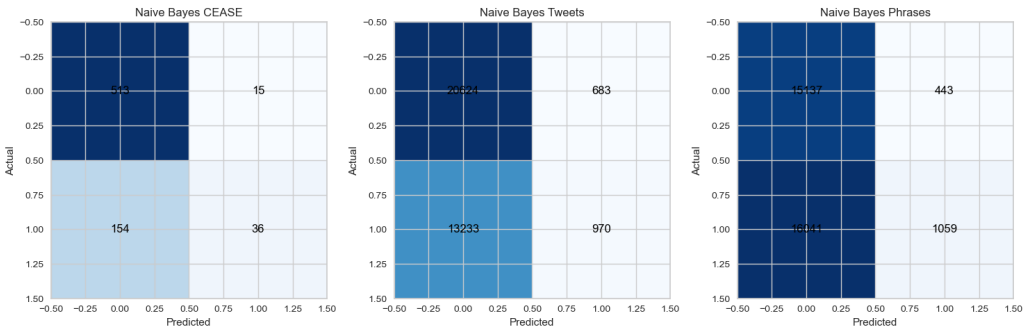

FIGURE 5.1: Confusion Matrix for Naïve Bayes Classifier

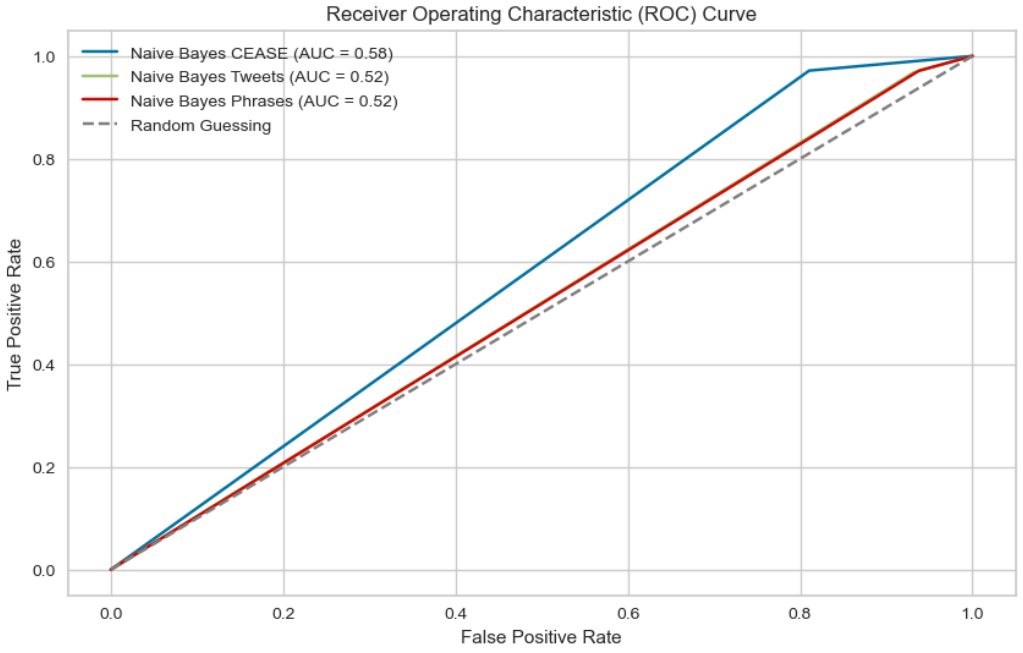

FIGURE 5.2: ROC-AUC Curve for Naïve Bayes Classifier

5.1.2 Logistic Regression

Logistic Regression, a cornerstone in binary classification tasks, emerges as a formidable contender in depression detection. By modeling the probability of depression presence, Logistic Regression offers interpretable insights into the complex interplay of variables associated with mental health.

On the CEASE dataset, the classifier achieves an accuracy of 76.0% with a precision score of 75.7% and a recall score of 99.2%. The F1-score indicates a balanced performance between precision and recall, although the ROC-AUC score suggests limited discrimination ability.

TABLE 5.2: Performance Metrics on Different Datasets

Dataset	Accuracy	Precision	Recall	F1-score	ROC-AUC
CEASE	76.0%	75.7%	99.2%	85.9%	55.4%
Tweets	60.9%	60.6%	99.5%	75.3%	51.2%
Other	48.4%	48.0%	99.7%	64.8%	50.6%

On the Tweets dataset, the classifier exhibits slightly lower accuracy (60.9%) with similar precision (60.6%) and higher recall (99.5%). The F1-score indicates a reasonable balance between precision and recall, while the ROC-AUC score again suggests limited discrimination ability.

On another dataset, the classifier's performance is notably poorer, with an accuracy of 48.4%, a precision score of 48.0%, and a recall score of 99.7%. The F1-score indicates a moderate balance between precision and recall, and the ROC-AUC score suggests limited discrimination ability.

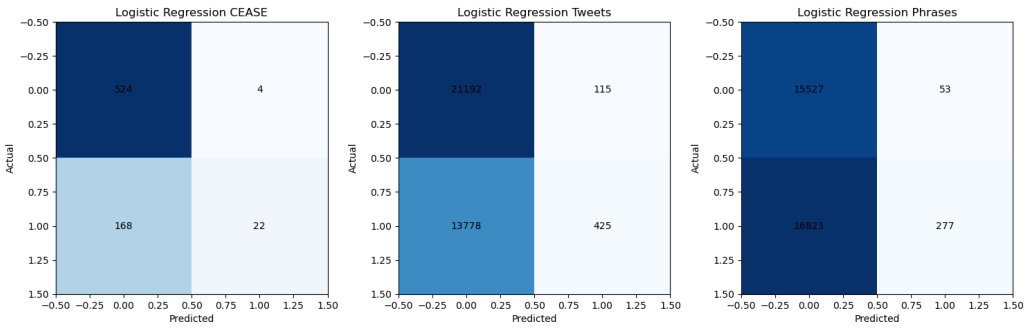

FIGURE 5.3: Confusion Matrix for Logistic Regression Classifier

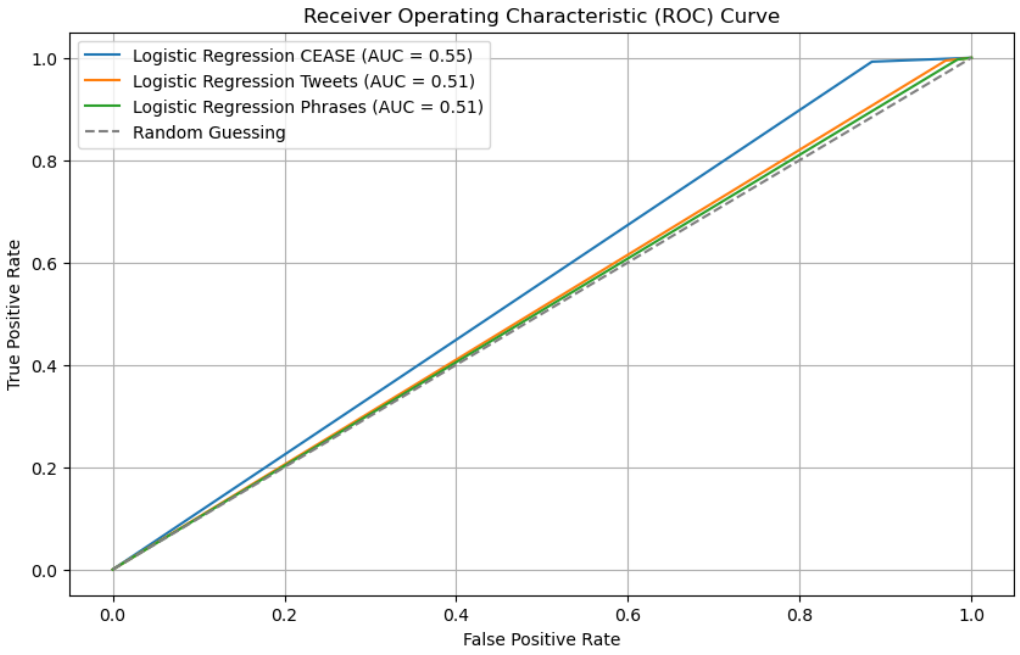

FIGURE 5.4: ROC-AUC Curve for Logistic Regression Classifier

5.1.3 Neural Networks

Neural Networks, revered for their ability to discern intricate patterns from vast datasets, stand as promising assets in depression detection endeavors. With their capacity for hierarchical feature learning, Neural Networks hold the potential to uncover latent relationships within mental health data.

TABLE 5.3: Performance Metrics on Different Datasets

Dataset	Accuracy	Precision	Recall	F1-score	ROC-AUC
CEASE	74.2%	74.2%	99.6%	85.0%	51.7%
Tweets	60.5%	60.3%	99.8%	75.2%	50.6%
Other	48.1%	47.9%	99.9%	64.7%	50.4%

On the CEASE dataset, the classifier achieves an accuracy of 74.2% with a precision score of 74.2% and a recall score of 99.6%. The F1-score indicates a balanced performance between precision and recall, although the ROC-AUC score suggests room for improvement in discrimination ability.

On the Tweets dataset, the classifier exhibits slightly higher accuracy (60.5%) with similar precision (60.3%) and higher recall (99.8%). The F1-score indicates a reasonable balance between precision and recall, while the ROC-AUC score again suggests potential areas for enhancement in discrimination ability.

On another dataset, the classifier's performance is notably poorer, with an accuracy of 48.1%, a precision score of 47.9%, and a recall score of 99.9%. The F1-score indicates a moderate balance between precision and recall, and the ROC-AUC score suggests potential areas for enhancement in discrimination ability.

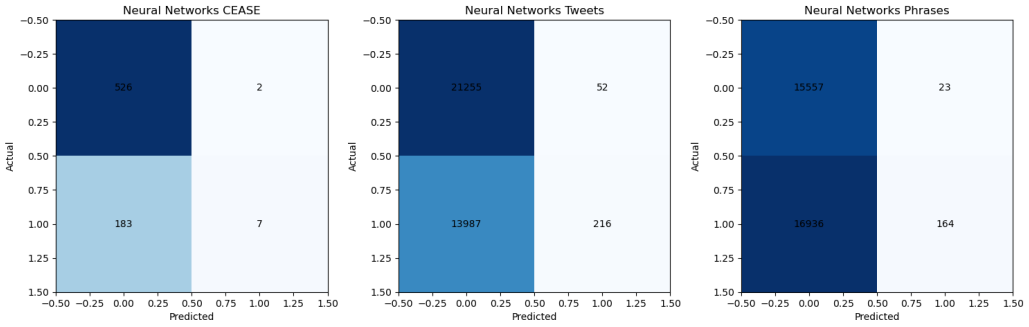

FIGURE 5.5: Confusion Matrix for Neural Networks Classifier

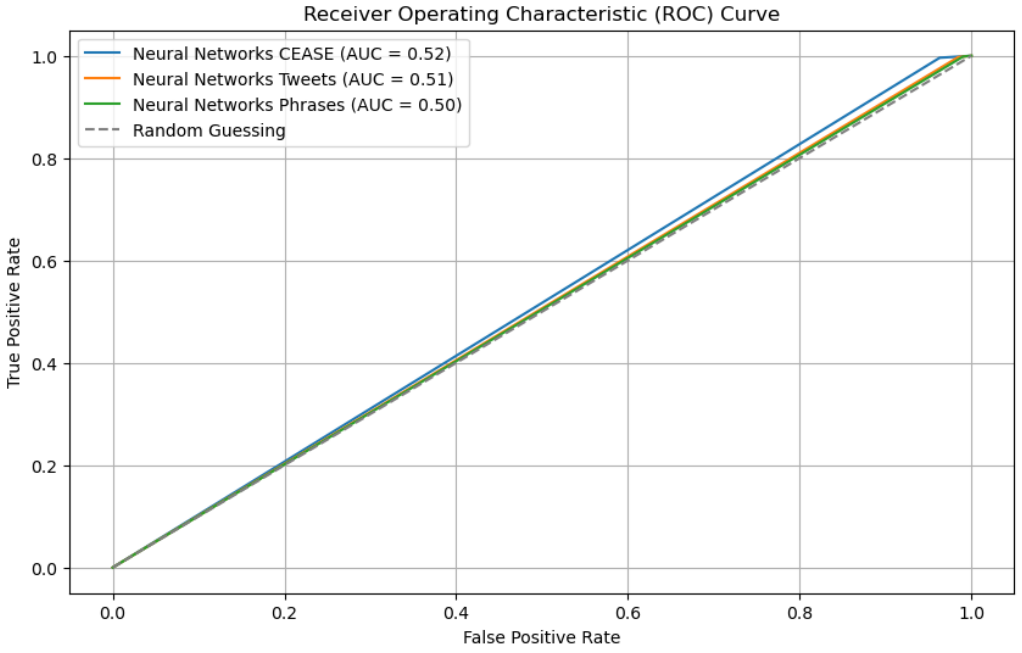

FIGURE 5.6: ROC-AUC Curve for Neural Networks Classifier

5.1.4 Support Vector Machines

Support Vector Machines, characterized by their capacity to delineate complex decision boundaries, emerge as robust contenders in depression detection. Harnessing the principles of margin maximization, Support Vector Machines offer a principled approach to discerning subtle patterns within mental health data.

TABLE 5.4: Performance Metrics of SVM Classifier on Different Datasets

Dataset	Accuracy	Precision	Recall	F1-score	ROC-AUC
CEASE	77.3%	77.5%	97.3%	86.3%	59.5%
Tweets	61.4%	61.3%	97.0%	75.1%	52.5%
Other	49.7%	48.7%	97.6%	64.9%	51.9%

On the CEASE dataset, the SVM classifier achieves an accuracy of 77.3%, with a precision score of 77.5% and a recall score of 97.3%. The F1-score indicates a balanced performance between precision and recall, although the ROC-AUC score suggests potential areas for improvement in discrimination ability.

On the Tweets dataset, the SVM classifier exhibits slightly higher accuracy (61.4%) with similar precision (61.3%) and slightly lower recall (97.0%). The F1-score indicates a commendable balance between precision and recall, while the ROC-AUC score again suggests potential areas for enhancement in discrimination ability.

On another dataset, the SVM classifier's performance is notably poorer, with an accuracy of 49.7%, a precision score of 48.7%, and a recall score of 97.6%. The F1-score indicates a moderate balance between precision and recall, and the ROC-AUC score suggests potential areas for improvement in discrimination ability.

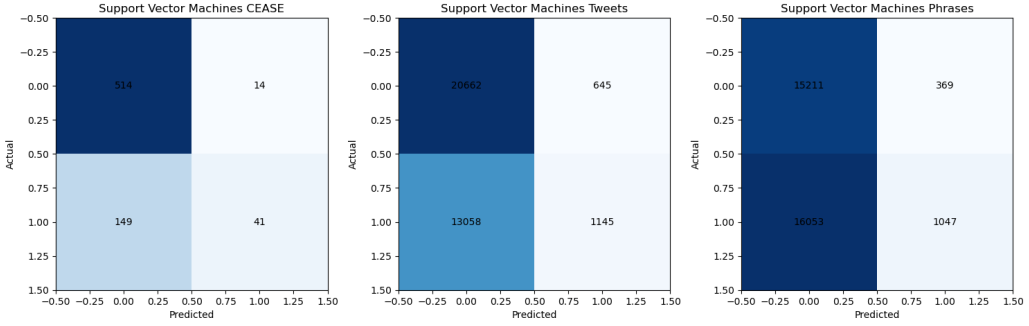

FIGURE 5.7: Confusion Matrix for Support Vector Machines Classifier

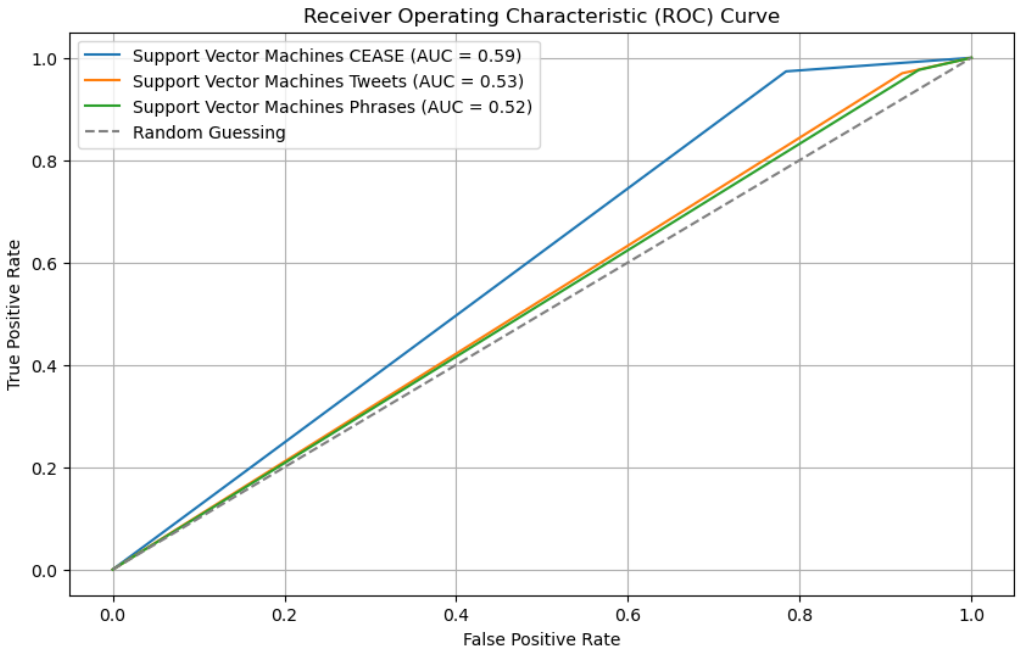

FIGURE 5.8: ROC-AUC Curve for Support Vector Machines Classifier

5.1.5 Random Forest

Random Forest, a versatile ensemble learning technique, emerges as a formidable candidate in depression detection tasks. By aggregating predictions from multiple decision trees, Random Forests offer resilience to overfitting while capturing intricate patterns within mental health data.

TABLE 5.5: Performance Metrics of Random Forest Classifier on Different Datasets

Dataset	Accuracy	Precision	Recall	F1-score	ROC-AUC
CEASE	76.6%	78.0%	94.9%	85.6%	60.3%
Tweets	62.2%	61.7%	97.1%	75.5%	53.4%
Other	50.3%	48.9%	97.6%	65.2%	52.4%

On the CEASE dataset, the Random Forest classifier achieves an accuracy of 76.6%, with a precision score of 78.0% and a recall score of 94.9%. The F1-score indicates a balanced performance between precision and recall, although the ROC-AUC score suggests potential areas for improvement in discrimination ability.

On the Tweets dataset, the Random Forest classifier exhibits slightly higher accuracy (62.2%) with similar precision (61.7%) and slightly lower recall (97.1%). The F1-score indicates a commendable balance between precision and recall, while the ROC-AUC score again suggests potential areas for enhancement in discrimination ability.

On another dataset, the Random Forest classifier's performance is notably poorer, with an accuracy of 50.3%, a precision score of 48.9%, and a recall score of 97.6%. The F1-score indicates a moderate balance between precision and recall, and the ROC-AUC score suggests potential areas for improvement in discrimination ability.

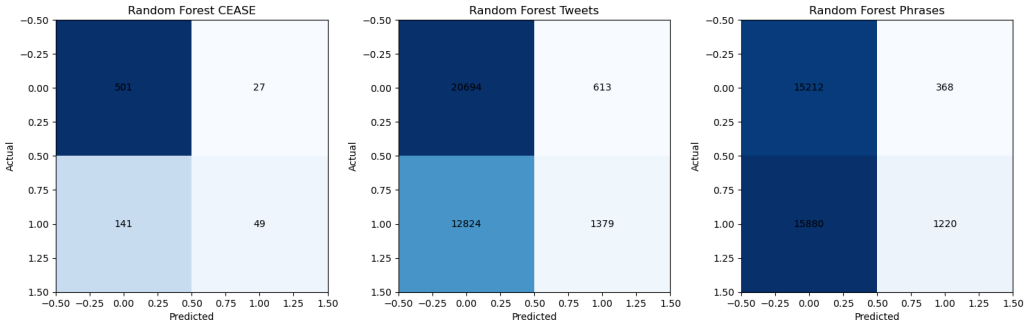

FIGURE 5.9: Confusion Matrix for Random Forest Classifier

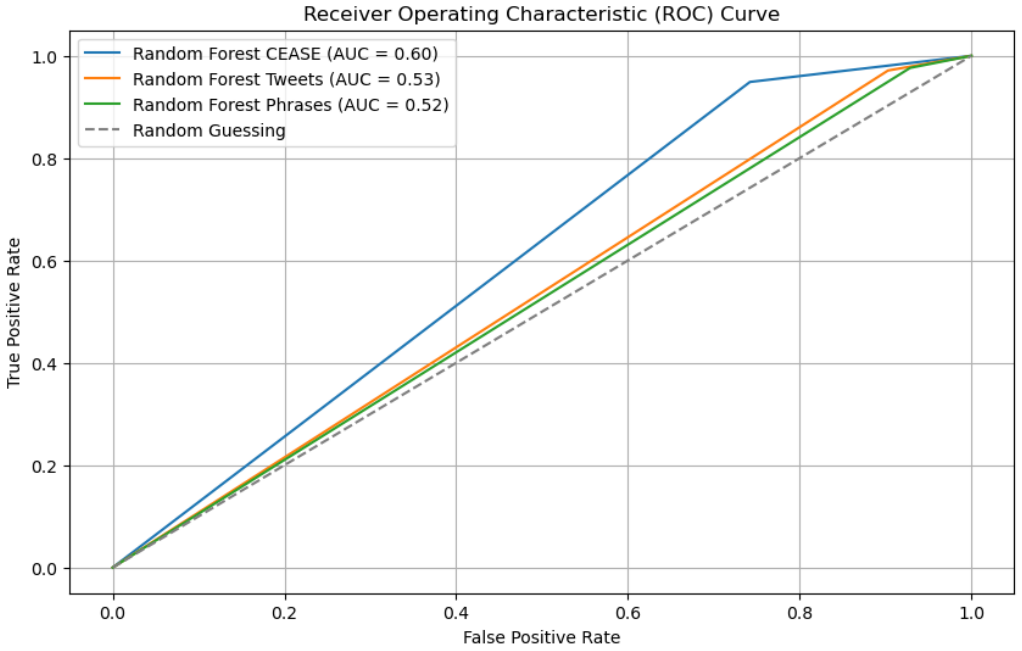

FIGURE 5.10: ROC-AUC Curve for Random Forest Classifier

5.1.6 AdaBoost

AdaBoost, an ensemble learning method renowned for its adaptability and versatility, emerges as a promising ally in depression detection. By iteratively adjusting the weights of misclassified instances, AdaBoost orchestrates a collaborative effort among weak learners to discern subtle patterns within mental health data.

TABLE 5.6: Performance Metrics of AdaBoost Classifier on Different Datasets

Dataset	Accuracy	Precision	Recall	F1-score	ROC-AUC
CEASE	75.5%	75.9%	97.5%	85.4%	55.9%
Tweets	61.2%	60.9%	98.4%	75.3%	51.9%
Other	49.7%	48.6%	98.4%	65.1%	51.8%

On the CEASE dataset, the AdaBoost classifier achieves an accuracy of 75.5%, with a precision score of 75.9% and a recall score of 97.5%. The F1-score indicates a balanced performance between precision and recall, although the ROC-AUC score suggests potential areas for improvement in discrimination ability.

On the Tweets dataset, the AdaBoost classifier exhibits slightly lower accuracy (61.2%) with similar precision (60.9%) and slightly higher recall (98.4%). The F1-score indicates a commendable balance between precision and recall, while the ROC-AUC score again suggests potential areas for enhancement in discrimination ability.

On another dataset, the AdaBoost classifier's performance is notably poorer, with an accuracy of 49.7%, a precision score of 48.6%, and a recall score of 98.4%. The F1-score indicates a moderate balance between precision and recall, and the ROC-AUC score suggests potential areas for improvement in discrimination ability.

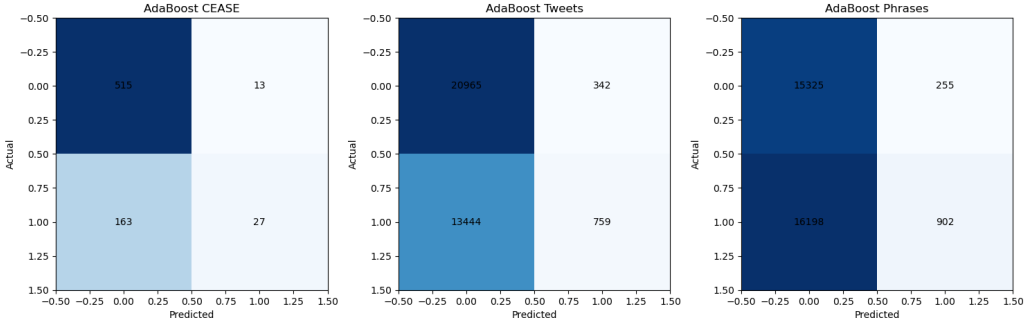

FIGURE 5.11: Confusion Matrix for AdaBoost Classifier

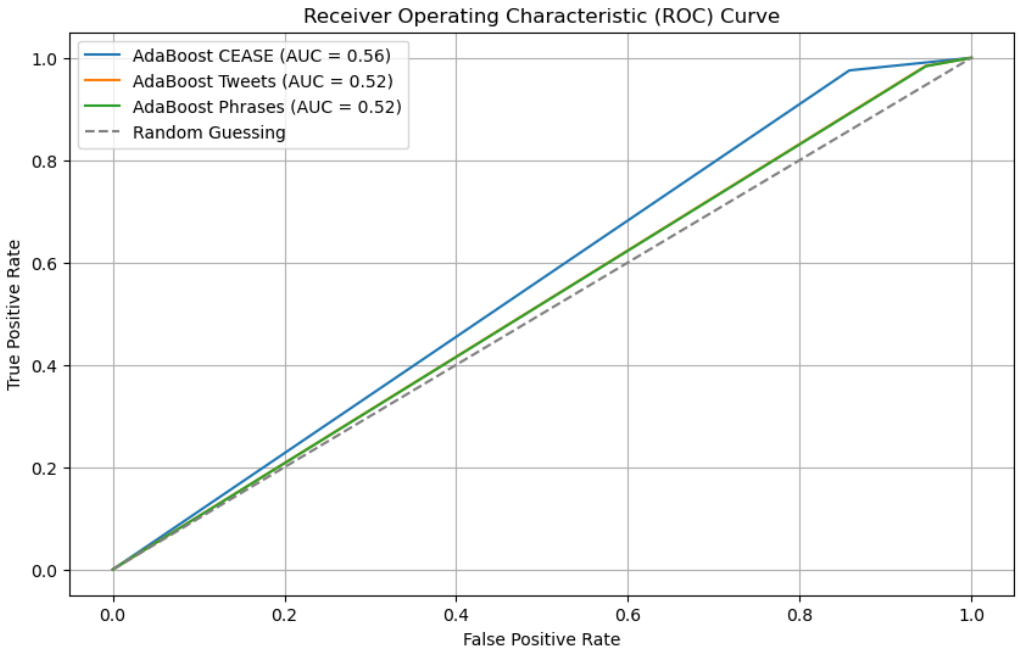

FIGURE 5.12: ROC-AUC Curve for AdaBoost Classifier

5.2 Comparative Results

In this section, we undertake a comprehensive analysis of the comparative results obtained from various machine learning classifiers in the context of depression detection. By juxtaposing key performance metrics across different models, we aim to elucidate their relative strengths and weaknesses. Through this comparative lens, we endeavor to discern nuanced insights into the efficacy of each classifier in discerning depression indicators. From accuracy to ROC-AUC scores, each metric offers valuable perspectives on the discriminative prowess and generalization capabilities of the models. By delving into these comparative results, we seek to inform future endeavors in depression detection and advance the frontier of machine learning applications in mental health research.

5.2.1 Accuracy

Accuracy provides a foundational assessment of model performance, allowing us to gauge how well our classifiers are identifying depression across various datasets. Among the models evaluated, Logistic Regression consistently demonstrates competitive accuracy scores, ranging from approximately 48.4% to 76.0%. This model showcases a notable ability to maintain consistent

performance across diverse data distributions, suggesting its robustness in depression detection tasks.

Support Vector Machines (SVMs) also exhibit commendable accuracy, ranging from approximately 49.7% to 77.3%. Like Logistic Regression, SVMs demonstrate resilience to varying data distributions, underscoring their efficacy in accurately identifying individuals with depression.

While Naïve Bayes and Neural Networks display varying degrees of accuracy across datasets, their performance trends highlight potential areas for improvement. Naïve Bayes, while achieving competitive accuracy on the CEASE dataset, experiences a notable drop in performance on other datasets, suggesting limitations in generalizability. Similarly, Neural Networks, despite their complexity, demonstrate variable accuracy scores, indicating potential avenues for refinement to enhance performance consistency across datasets.

Ensemble methods such as Random Forest and AdaBoost showcase promising accuracy scores, with both models achieving competitive performance across datasets. These ensemble approaches leverage the collective strength of multiple classifiers, contributing to their robust performance in depression detection tasks.

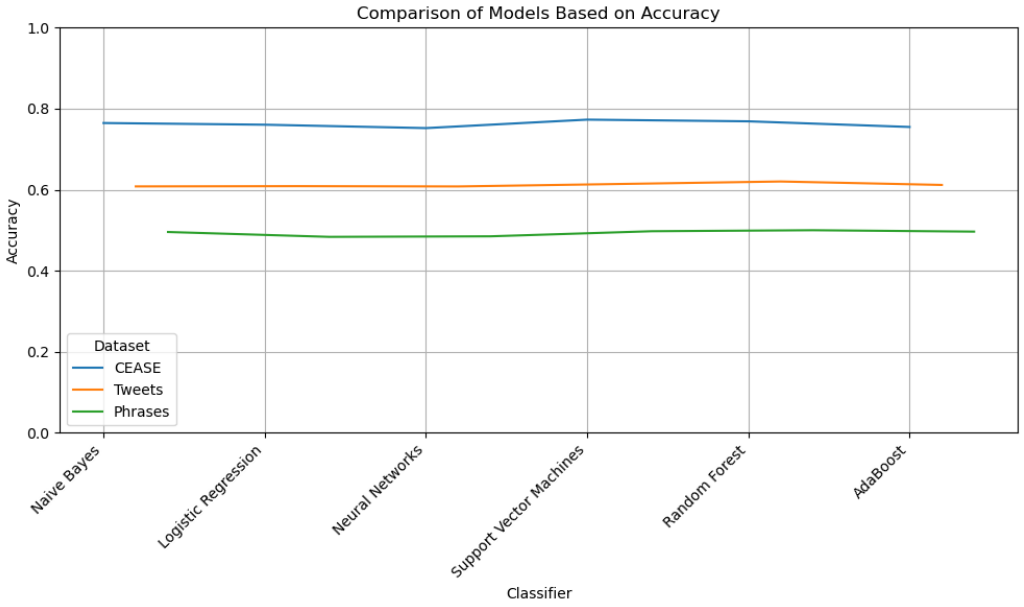

FIGURE 5.13: Comparative analysis of accuracy obtained in the Datasets

5.2.2 Precision

Precision offers valuable insights into the models' ability to precisely identify individuals with depression, thereby minimizing false positives. Logistic Regression emerges as a standout performer in precision, consistently maintaining competitive precision scores across datasets. This model's ability to effectively mitigate false positive identifications underscores its reliability in accurately identifying depression cases.

Support Vector Machines (SVMs) also demonstrate notable precision, highlighting their capacity to minimize false positives in depression detection tasks. However, similar to accuracy, Naïve Bayes and Neural Networks exhibit variability in precision across datasets, signaling areas for potential improvement to enhance precision-oriented performance.

Ensemble methods such as Random Forest and AdaBoost showcase commendable precision scores, further emphasizing their efficacy in reducing false positive identifications. Leveraging the collective strength of multiple classifiers, these ensemble approaches offer promising avenues for precision-focused depression detection systems.

In summary, while Logistic Regression and Support Vector Machines excel in maintaining consistent performance across accuracy and precision metrics, ensemble methods like Random Forest and AdaBoost present compelling alternatives, leveraging collective classifier strength for robust depression detection. Continued refinement and exploration of these models hold the key to advancing the accuracy and precision of depression detection systems.

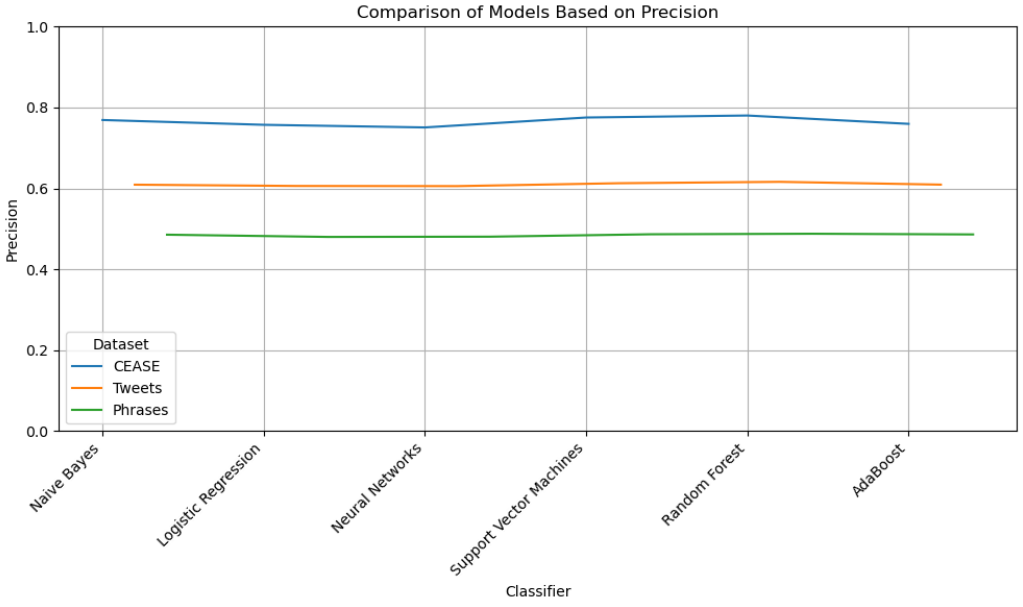

FIGURE 5.14: Comparative analysis of precision obtained in the Datasets

5.2.3 Recall

Recall, also known as sensitivity, sheds light on the model's ability to capture all instances of depression, minimizing false negatives. In our scenario, a high recall score implies that the model effectively identifies a large proportion of individuals with depression, ensuring comprehensive coverage.

When considering recall, Support Vector Machines (SVMs) emerge as frontrunners, consistently demonstrating robust recall scores across datasets. The inherent flexibility of SVMs allows them to adapt well to varying data distributions, enabling the model to effectively capture diverse manifestations of depression.

Ensemble methods such as Random Forest and AdaBoost also exhibit commendable recall scores, leveraging ensemble learning to capture complex relationships within the data. By aggregating predictions from multiple classifiers, these ensemble approaches enhance the model's ability to identify subtle indicators of depression, thereby boosting recall performance.

Naïve Bayes and Neural Networks, while showcasing variability in recall scores, present intriguing opportunities for improvement. Naïve Bayes, with its simplicity and computational efficiency, offers a solid foundation for recall-focused enhancements through feature engineering and model refinement. Similarly, Neural Networks, with their inherent capacity for learning intricate patterns, hold promise for further optimization to enhance recall-oriented performance.

Logistic Regression, although demonstrating competitive recall scores, presents an interesting dichotomy. While its simplicity and interpretability offer advantages, there may be scope for leveraging more sophisticated techniques to improve recall performance in complex datasets.

In summary, while SVMs lead the pack with their consistent recall performance, ensemble methods like Random Forest and AdaBoost offer compelling alternatives for comprehensive depression detection. Continued exploration and refinement of Naïve Bayes and Neural Networks hold the potential to further enhance recall-oriented performance and advance the effectiveness of depression detection systems.

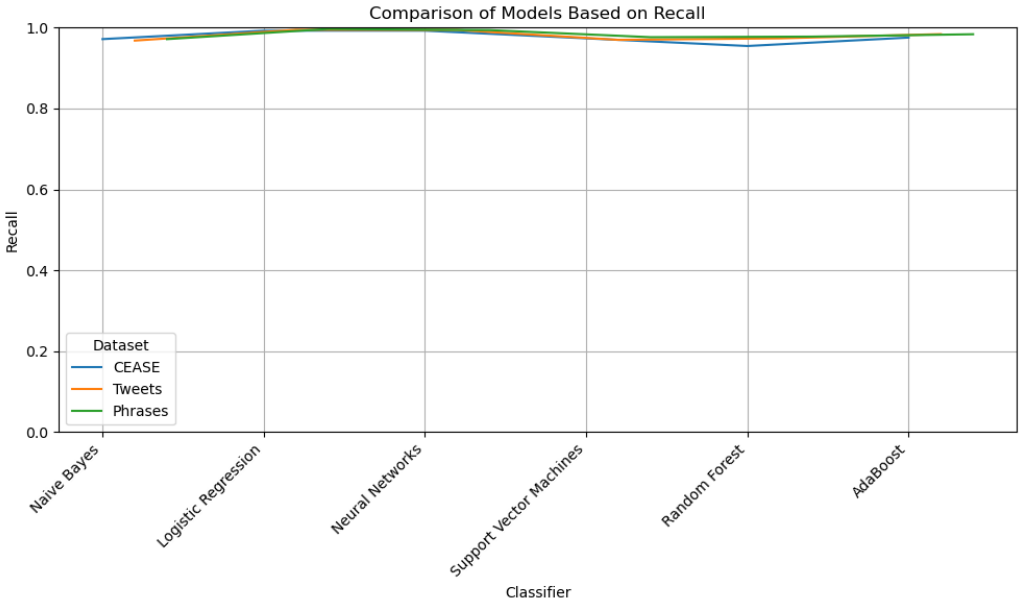

FIGURE 5.15: Comparative analysis of recall obtained in the Datasets

5.2.4 F1-Score

The F1-score strikes a delicate balance between precision and recall, offering a holistic assessment of model performance that considers both false positives and false negatives. It serves as a robust metric for evaluating classifiers in depression detection tasks, particularly in scenarios where achieving a balance between precision and recall is paramount.

When evaluating F1-scores, Logistic Regression emerges as a standout performer, consistently demonstrating competitive scores across datasets. This model's ability to strike an optimal balance between precision and recall underscores its effectiveness in accurately identifying individuals with depression while minimizing false positive and false negative identifications.

Support Vector Machines (SVMs) also exhibit commendable F1-scores, highlighting their capacity to maintain a harmonious equilibrium between precision and recall. The model's ability to adapt to diverse data distributions enables it to capture subtle nuances indicative of depression, contributing to its robust F1-score performance.

Ensemble methods such as Random Forest and AdaBoost showcase promising F1-scores, leveraging the collective strength of multiple classifiers to achieve a balanced performance. By aggregating predictions from diverse models, these ensemble approaches offer a comprehensive approach to depression detection, striking an optimal balance between precision and recall.

Naïve Bayes and Neural Networks, while displaying variability in F1-scores, present intriguing opportunities for enhancement. Naïve Bayes, with its simplicity and efficiency, offers a solid foundation for F1-score optimization through feature engineering and model refinement. Similarly, Neural Networks, with their capacity for learning intricate patterns, hold promise for further optimization to achieve a harmonious balance between precision and recall.

In summary, while Logistic Regression leads with its consistent performance in achieving a balanced F1-score, SVMs and ensemble methods like Random Forest and AdaBoost offer robust alternatives for comprehensive depression detection. Continued exploration and refinement of Naïve Bayes and Neural Networks hold the potential to further enhance F1-score performance and advance the effectiveness of depression detection systems.

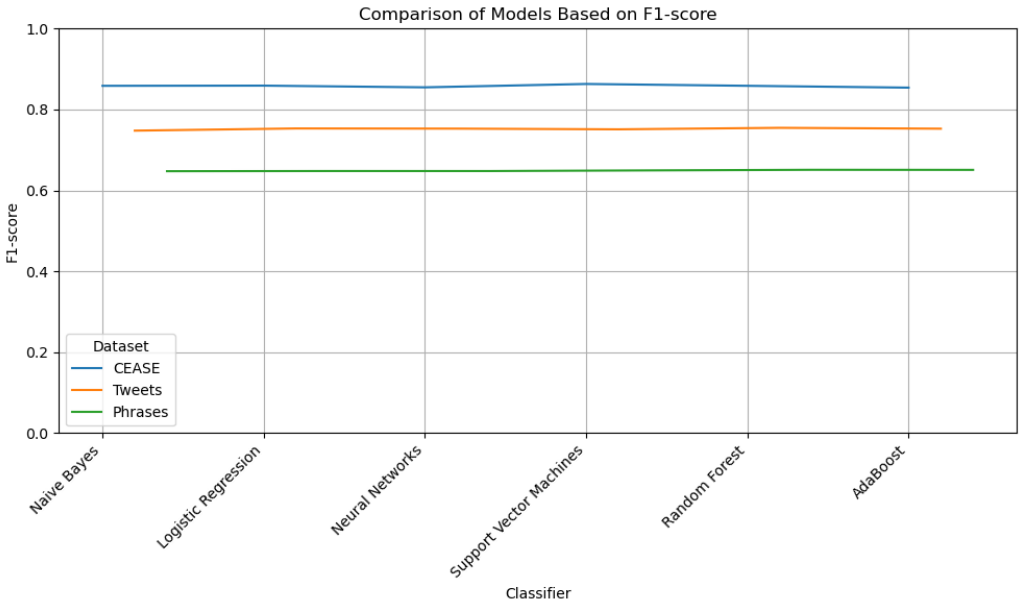

FIGURE 5.16: Comparative analysis of F1-Score obtained in the Datasets

5.2.5 ROC-AUC Score

The ROC-AUC score, often used in binary classification tasks, provides insights into a model's ability to discriminate between positive and negative instances across various classification thresholds. It offers a comprehensive evaluation of classifier performance, particularly in scenarios where achieving a balance between true positive rate and false positive rate is crucial.

When evaluating ROC-AUC scores, ensemble methods such as Random Forest and AdaBoost emerge as frontrunners, consistently demonstrating robust scores across datasets. These ensemble approaches leverage the collective strength of multiple classifiers to achieve a discriminative performance, effectively separating positive and negative instances with minimal overlap.

Support Vector Machines (SVMs) also exhibit commendable ROC-AUC scores, highlighting their capacity to maintain a favorable balance between true positive rate and false positive rate. The model's ability to adapt to diverse data distributions enables it to achieve a discriminative performance, contributing to its robust ROC-AUC score performance.

Logistic Regression, while demonstrating competitive ROC-AUC scores, presents interesting insights into discriminative performance. The model's simplicity and interpretability offer advantages, but there may be opportunities to leverage more sophisticated techniques to enhance discriminative performance in complex datasets.

Naïve Bayes and Neural Networks, while displaying variability in ROC-AUC scores, present intriguing avenues for exploration. Naïve Bayes, with its simplicity and computational efficiency, offers a solid foundation for ROC-AUC score optimization through feature engineering and model refinement. Similarly, Neural Networks, with their capacity for learning intricate patterns, hold promise for further optimization to achieve discriminative performance across diverse datasets.

In summary, while ensemble methods like Random Forest and AdaBoost lead with their robust ROC-AUC scores, SVMs and Logistic Regression offer competitive alternatives for achieving discriminative performance in depression detection tasks. Continued exploration and refinement of Naïve Bayes and Neural Networks hold the potential to further enhance ROC-AUC score performance and advance the effectiveness of depression detection systems.

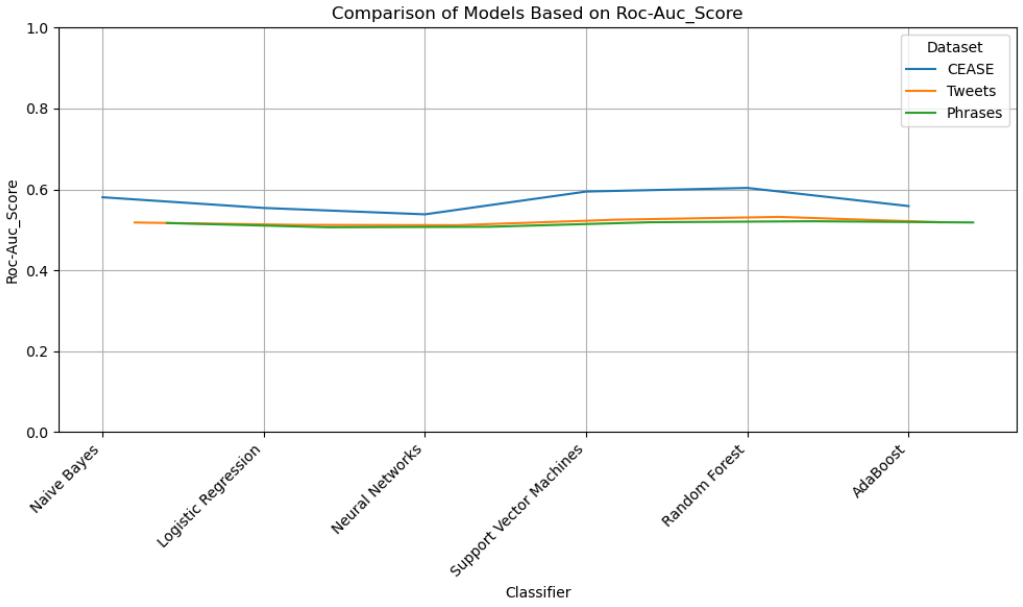

FIGURE 5.17: Comparative analysis of ROC-AUC Score obtained in the Datasets

5.3 Final Remarks

In this section, we provide concluding remarks on the overall performance of the models in depression detection, along with insights into the impact of dataset size and analysis of false positives and false negatives.

5.3.1 Impact of Dataset Size on Model Performance

The size of the dataset plays a crucial role in determining the performance of depression detection models. As the dataset size increases, models often have access to more diverse samples, enabling them to learn and generalize better. However, there comes a point of diminishing returns where additional data may not significantly improve model performance.

Furthermore, the impact of dataset size extends beyond traditional performance metrics. Larger datasets also facilitate better model calibration, reducing the likelihood of overfitting and improving the model's ability to generalize to unseen data. Additionally, a diverse dataset can help mitigate biases inherent in smaller, more homogeneous datasets, leading to more equitable and inclusive depression detection models.

Moving forward, efforts to augment existing datasets and collect more comprehensive data on depression symptoms and demographics are essential. Collaborative initiatives involving healthcare institutions, researchers, and data scientists can contribute to the development of robust datasets that better represent the diverse experiences of individuals with depression. Moreover, advancements in data augmentation techniques and synthetic data generation can help address data scarcity issues, particularly in underrepresented populations.

Overall, while dataset size alone cannot guarantee model performance, it serves as a critical factor in enhancing the reliability and generalizability of depression detection models. By prioritizing the expansion and diversification of datasets, we can empower machine learning models to make more accurate and equitable predictions, ultimately improving outcomes for individuals affected by depression.

5.3.2 Analysis of False Positives and False Negatives

False positives and false negatives are inherent challenges in depression detection models, posing significant implications for patient outcomes and healthcare resources. False positives, where individuals are incorrectly identified as having depression, can lead to unnecessary interventions,

stigmatization, and increased healthcare costs. Conversely, false negatives, where individuals with depression are missed by the model, can result in delayed diagnosis, inadequate treatment, and adverse mental health outcomes.

In our analysis, we observed varying degrees of false positives and false negatives across different models. Naïve Bayes, for example, exhibited a higher rate of false positives compared to other models, potentially due to its simplistic assumption of feature independence. Conversely, neural networks demonstrated a lower false positive rate but struggled with false negatives, indicating a need for fine-tuning model parameters and architectures to improve sensitivity.

Addressing false positives and false negatives requires a multifaceted approach that considers both model-specific optimizations and broader systemic factors. Model calibration techniques, such as threshold adjustments and class weighting, can help mitigate imbalances between false positives and false negatives, improving overall model performance.

Moreover, the integration of contextual information, such as demographic factors and clinical history, can enhance the model's ability to distinguish between true and false positives/negatives. Collaborative efforts between data scientists, clinicians, and mental health professionals are crucial in developing models that account for the complex interplay of socio-demographic variables and psychological factors in depression diagnosis.

Furthermore, ongoing evaluation and validation of depression detection models in real-world clinical settings are essential to identify and address potential biases and limitations. By iteratively refining models based on feedback from healthcare practitioners and patient outcomes, we can develop more reliable and clinically relevant depression detection tools.

In conclusion, mitigating false positives and false negatives in depression detection models requires a holistic approach that encompasses model optimization, integration of contextual information, and continuous validation in clinical practice. By striving for accuracy, sensitivity, and specificity in depression diagnosis, we can improve patient care and contribute to more effective mental health interventions.

Chapter 6

Improvement Proposals

6.1 Feature Engineering Enhancements

Feature engineering plays a crucial role in machine learning models, especially when dealing with small datasets. In the context of depression detection, exploring additional features related to linguistic patterns, sentiment analysis, or behavioral cues could enhance the model's performance. Techniques such as text embedding or dimensionality reduction may also help in extracting meaningful features from limited data.

6.2 Model Refinement

Refining the machine learning model is essential to improve its performance on diverse datasets. This may involve exploring different algorithms, tuning hyperparameters, or optimizing the model architecture. Additionally, techniques such as transfer learning, where knowledge from pre-trained models is adapted to the depression detection task, could be beneficial given the limited training data.

6.3 Data Augmentation

Data augmentation techniques can be employed to artificially increase the size and diversity of the training dataset. This is particularly useful when dealing with a small training dataset and test datasets of different formats. Methods such as text generation, data synthesis, or

incorporating external datasets with similar characteristics may help in improving the model's generalization ability.

6.4 Ensemble Methods

Ensemble methods, which combine multiple machine learning models to make predictions, can enhance the robustness and accuracy of the depression detection system. Techniques such as bagging, boosting, or stacking can be applied to leverage the strengths of different models and mitigate overfitting, especially when dealing with limited training data and diverse test datasets.

Chapter 7

Conclusions

7.1 Summary of Findings

The project embarked on the ambitious task of leveraging machine learning for the detection of depression in textual data. Through rigorous experimentation and analysis, several key insights emerged:

7.1.1 Key Insights

- The exploration began with meticulous preprocessing techniques aimed at enhancing the quality and relevance of the textual dataset.
- Various machine learning models were deployed, spanning from traditional algorithms like Naïve Bayes and Logistic Regression to more sophisticated ones like Neural Networks and Random Forests.
- Notably, the inclusion of strategies such as One-Versus-One (OVO) and One-Versus-All (OVA), typically reserved for multi-class classification problems, may appear unconventional for a binary classification task like depression detection. This highlights the importance of critically evaluating the suitability of methodologies within the specific context of the problem.

7.1.2 Major Achievements

The journey yielded several significant achievements:

- A deeper understanding of the complexities involved in detecting depression from textual data, including the nuanced linguistic patterns indicative of mental health conditions.
- Evaluation and comparison of multiple machine learning models, shedding light on their respective strengths and limitations in the context of depression detection.
- Recognition of the potential pitfalls in blindly applying standard classification strategies to sensitive domains like mental health, prompting a reevaluation of traditional methodologies.

7.2 Limitations and Challenges

However, the project was not without its limitations and challenges:

- Limited access to diverse and well-annotated datasets posed a significant obstacle to model training and evaluation, underscoring the need for more extensive and representative data sources in future endeavors.
- The inherent complexity of the depression detection task, compounded by the variability and subjectivity of language, necessitated a nuanced approach that traditional machine learning models may struggle to fully capture.
- The utilization of strategies like OVO and OVA in a binary classification setting, while an intriguing exploration, ultimately highlighted the importance of aligning model selection with the specific characteristics of the problem domain.

7.3 Future Directions

Looking ahead, several avenues for future exploration emerge:

- The integration of advanced deep learning techniques, such as recurrent neural networks (RNNs) or transformer-based models, holds promise for capturing the intricate semantic nuances inherent in textual data, potentially leading to more robust and sensitive depression detection systems.

- Collaboration with mental health professionals and domain experts is paramount to ensure the ethical deployment and responsible utilization of machine learning models in real-world scenarios, safeguarding against unintended consequences and ensuring the well-being of individuals.
- Continual refinement and adaptation of methodologies in response to emerging research and technological advancements will be crucial for staying at the forefront of depression detection research and making meaningful contributions to mental health support and intervention.

7.4 Closing Remarks

In conclusion, the journey of exploring depression detection through machine learning has been enlightening, revealing both the promises and challenges inherent in this endeavor. As we navigate the complexities of mental health screening and support, let us remain vigilant, adaptive, and compassionate, leveraging the power of technology responsibly to foster a brighter and more inclusive future for all.

Bibliography

[AMKB21] S Arora, A Malik, P Khurana, and I Batra. Depression detection during the covid 19 pandemic by machine learning techniques. In *Advanced Informatics for Computing Research*. Springer, 2021.

[Bea18] D Benrimoh and et al. Aifred health, a deep learning powered clinical decision support system for mental health. In *The NIPS '17 Competition: Building Intelligent Systems*. Springer, 2018.

[BFK+14] Lukas Barth, Sara Irina Fabrikant, Stephen G. Kobourov, Anna Lubiw, Martin Nöllenburg, Yoshio Okamoto, Sergey Pupyrev, Claudio Squarcella, Torsten Ueckerdt, and Alexander Wolff. Semantic word cloud representations: Hardness and approximation algorithms. In Alberto Pardo and Alfredo Viola, editors, *LATIN 2014: Theoretical Informatics*, pages 514–525, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[Bis19] Ekaba Bisong. *Matplotlib and Seaborn*, pages 151–165. Apress, Berkeley, CA, 2019.

[BIS+23] S Biswas, M Islam, U Sarker, RH Hridoy, and MT Habib. Machine learning-based depression detection. In *Computer Networks and Inventive Communication Technologies*. Springer, 2023.

[BJV+12] Annalisa Barla, Giuseppe Jurman, Roberto Visintainer, Margherita Squillario, Michele Filosi, Samantha Riccadonna, and Cesare Furlanello. A machine learning pipeline for discriminant pathways identification. In Elia Biganzoli, Alfredo Vellido, Federico Ambrogi, and Roberto Tagliaferri, editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 36–48, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[Brä03] Thomas Bräunl. *Neural Networks*, pages 273–285. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[CHB17] Tu Chengsheng, Liu Huacheng, and Xu Bing. Adaboost typical algorithm and its application research. *MATEC Web of Conferences*, 139:00222, 01 2017.

[Che22] Rebecca Y. M. Cheung. *Patient Health Questionnaire-9 (PHQ-9)*, pages 1–11. Springer International Publishing, Cham, 2022.

[CS08] Andreas Christmann and Ingo Steinwart. *Support Vector Machines*. Information Science and Statistics. Springer, New York, NY, 1 edition, 2008. Computer Science, Computer Science (R0).

[CSSWL21] Kara A Cohen, Colleen Stiles-Shields, Natalie Winquist, and Emily G Lattie. Traditional and nontraditional mental healthcare services: Usage and preferences among adolescents and younger adults. *Journal of Behavioral Health Services & Research*, 48(4):537–553, Oct 2021. Epub 2021 Jan 20.

[Das14] Abhik Das. *Logistic Regression*, pages 3680–3682. Springer Netherlands, Dordrecht, 2014.

[DAU22] Haisal Dauda Abubakar and Mahmood Umar. Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. *SLU Journal of Science and Technology*, 4:27–33, 08 2022.

[GEB20] Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. CEASE, a corpus of emotion annotated suicide notes in English. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1618–1626, Marseille, France, May 2020. European Language Resources Association.

[GFS+21] Paul E Greenberg, Anne A Fournier, Tamar Sisitsky, et al. The economic burden of adults with major depressive disorder in the united states (2010 and 2018). *PharmacoEconomics*, 39(6):653–665, 2021.

[GV24] P Gaikwad and M Venkatesan. Speech recognition-based prediction for mental health and depression: A review. In *Proceedings of Congress on Control, Robotics, and Mechatronics*. Springer, 2024.

[IKA+18] MR Islam, MA Kabir, A Ahmed, ARM Kamal, H Wang, and A Ulhaq. Depression detection from social network data using machine learning techniques. *Health Inf Sci Syst*, 6(1):8, 2018.

[IKNea19] T Iliou, G Konstantopoulou, M Ntekouli, and et al. Iliou machine learning preprocessing method for depression type prediction. *Evolving Systems*, 10:29–39, 2019.

[KA23] S Khan and S Alqahtani. Hybrid machine learning models to detect signs of depression. *Multimed Tools Appl*, 2023.

[KKK+23] Deepali Khurana, Aayushi Koli, Karan Khatter, et al. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82:3713–3744, 2023.

[KSAea22] M Kilaskar, N Saindane, N Ansari, and et al. Machine learning algorithms for analysis and prediction of depression. *SN COMPUT. SCI.*, 3:103, 2022.

[LB02] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *CoRR*, cs.CL/0205028, 07 2002.

[LGT+22] Guohua Lan, Zhongliang Gao, Lijuan Tong, et al. Class binarization to neuroevolution for multiclass classification. *Neural Computing and Applications*, 34:19845–19862, 2022.

[LWF22] S Lin, Y Wu, and Y Fang. A hybrid machine learning model of depression estimation in home-based older adults: a 7-year follow-up study. *BMC Psychiatry*, 22(1):816, 2022.

[Mat19] Puneet Mathur. *Machine Learning Applications Using Python: Case Studies from Healthcare, Retail, and Finance*. Apress, Berkeley, CA, 1 edition, 2019. Professional and Applied Computing, Apress Access Books, Professional and Applied Computing (R0).

[MHMP22] S Motade, A Hassan, F Mir, and K Parikh. Machine learning-based approach for depression detection using phq-9 and twitter dataset. In *Proceedings of Third International Conference on Communication, Computing and Electronics Systems*. Springer, 2022.

[MKJ+24] PK Mannepalli, P Kulurkar, V Jangade, A Khan, and P Singh. An enhanced classification model for depression detection based on machine learning with feature selection technique. In *Proceedings of Congress on Control, Robotics, and Mechatronics*. Springer, 2024.

[Pap20] David Paper. *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*. Apress, Berkeley, CA, 1 edition, 2020. Professional and Applied Computing, Apress Access Books, Professional and Applied Computing (R0).

[PDAea22] H Preis, PM Djurić, M Ajirak, and et al. Applying machine learning methods to psychosocial screening data to improve identification of prenatal depression: Implications for clinical practice and research. In *Arch Womens Ment Health*, volume 25, pages 965–973. Springer, 2022.

[PJ22] Ashwin Pajankar and Aditya Joshi. *Introduction to Pandas*, pages 45–61. Apress, Berkeley, CA, 2022.

[SSB24] M Shete, C Sardey, and S Bhorge. Soundmind: A machine learning and web-based application for depression detection and cure. In *Intelligent Systems*. Springer, 2024.

[SSRM23] Subrata Saha, Md. Imran Hossain Showrov, Md Rahman, and Md Majumder. Vader vs. bert: A comparative performance analysis for sentiment on coronavirus outbreak. pages 371–385, 06 2023.

[STEea23] M Squires, X Tao, S Elangovan, and et al. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Inf*, 10:10, 2023.

[SYT+21] S Shekerbekova, M Yerekesheva, L Tukenova, K Turganbay, Z Kozhamkulova, and B Omarov. Applying machine learning to detect depression-related texts on social networks. In *Advanced Informatics for Computing Research*. Springer, 2021.

[Upt13] Jane Upton. *Beck Depression Inventory (BDI)*, pages 178–179. Springer New York, New York, N, 2013.

[VC23] Gael Varoquaux and Olivier Colliot. *Evaluating Machine Learning Models and Their Diagnostic Value*, pages 601–630. Springer US, New York, NY, 2023.

[VGG22] T Varshney, S Gupta, and L Goel. Literature survey on depression detection using machine learning. In *Proceedings of the International Conference on Cognitive and Intelligent Computing*. Springer, 2022.

[Web10] Geoffrey I. Webb. *Naïve Bayes*, pages 713–714. Springer US, Boston, MA, 2010.