UNIVERSITY OF EMBU

IAN IGADO MMBAYI

MSc

2025

SENTIMENT ANALYSIS-BASED MODEL FOR MONITORING USER ENGAGEMENT WITH MENTAL HEALTH CHATBOTS

IAN IGADO MMBAYI (MSc)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION TECHNOLOGY IN THE UNIVERSITY OF EMBU

NOVEMBER, 2025

DECLARATION

This thesis is my original work and has not been presented for a degree in any other university.

Signature……..... Date…….....…….

Ian Igado Mmbayi

Department of Computing and Information Technology

B532/1648/2022

This thesis has been submitted for examination with our approval as the University supervisors.

Signature……..... Date…….....…….

Dr. Consolata Gakii

Department of Computing and Information Technology

University of Embu

Signature……..... Date…….....…….

Dr. Faith Mueni Musyoka

Department of Computing and Information Technology

University of Embu

ACKNOWLEDGEMENT

First and foremost, I am deeply grateful to God for granting me strength, patience, and perseverance throughout this journey. I would like to express my sincere gratitude to my supervisors, Dr. Consolata Gakii and Dr. Faith Mueni Musyoka, for their invaluable guidance, support, and constructive feedback at every stage of this research. Their expertise and encouragement have been instrumental in shaping the direction and quality of my work. I am also thankful to the academic and administrative staff of the Department of Computing and Information Technology at the University of Embu for providing a conducive environment and resources necessary for the completion of this study. Special thanks go to my family for their unwavering love and encouragement, especially during moments when the journey felt overwhelming. Your belief in me has been my greatest source of motivation. To my friends and colleagues, thank you for the stimulating discussions, for listening when I needed to vent, and for being a constant source of support and laughter. Lastly, I am grateful to all the individuals and institutions who contributed in any way to the successful completion of this thesis. Your support has made this milestone possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABSA	Aspect-Based Sentiment Analysis
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CBT	Cognitive Behaviour Therapy
CNN	Convolutional Neural Networks
IDF	Inverse Document Frequency
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MNB	Multinomial Naive Bayes
NB	Naïve Bayes
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
PTSD	Post-Traumatic Stress Disorder
RF	Random Forest
RNNs	Recurrent Neural Networks
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machines
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency

ABSTRACT

Mental health challenges, particularly among youth, are compounded by stigma and limited access to professional care. This has driven demand for scalable digital solutions like chatbots. This study introduces a sentiment analysis-based model to assess user satisfaction with mental health chatbots, analysing 82,102 reviews from six popular apps on Google Play and Apple's App Stores. A multi-class sentiment classification of positive, negative, and neutral was implemented, enhanced by Synthetic Minority Over-sampling Technique for class balancing, comparing five traditional machine learning models with Bidirectional Encoder Representations from Transformers, a transformer model. Random Forest achieved 98.18% accuracy among traditional models, while BERT outperformed all with 99.17% accuracy, surpassing prior benchmarks. Aspect-based analysis revealed that Emotion and Usability drive positive feedback, while Reliability issues fuel negative sentiments, offering actionable insights for developers to enhance chatbot design. This work advances digital mental health research by integrating multi-class classification, transformer models, and aspect-based analysis, demonstrating a scalable framework for evaluating user feedback.

Keywords: Sentiment Analysis, Mental Health, Chatbots, BERT, User Reviews, Machine Learning, SMOTE, Aspect-Based Analysis

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Severe mental health conditions such as depression, anxiety, and stress have reached alarming heights globally, with more than 300 million affected and with far-reaching personal, social, and economic impacts (WHO, 2022). Apart from causing significant disability, these conditions also produce economic losses of approximately $1 trillion annually through decreased productivity, added healthcare costs, and long-term social support (Chisholm et al., 2016). The scenario is even worse in low- and middle-income countries such as Kenya, where over 75% of the affected with mental health comorbidities never receive the necessary care. This largely stems from the absence of mental health professionals, as only 0.19 psychiatrists per 100,000 population with few facilities and deep-rooted cultural stigma deterring most to seek care (Mutiso et al., 2021). The situation is further exacerbated by the allocation of less than 1% of the national healthcare allocation in the Kenyan national health budget for mental health (Ministry of Health, Kenya, 2021). The COVID-19 pandemic further worsened the situation, occasioning global anxiety and depression upsurges of over 25%, fueled by isolation, financial stress, and decreased care accessibility (Shan et al., 2022).

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines and enables them to perform tasks such as learning, problem solving, and decision making (Bharadiya, 2022), it is a field that is rapidly growing and gaining importance in today's world. In business, it allows businesses to analyze big data and provide insights, aid decision-making, and improve business strategies (Vamathevan et al., 2019). AI helps optimize the manufacturing process in Micro-Electro-Mechanical Systems based sensors by increasing efficiency and accuracy (Podder et al., 2023). It is widely used for recognizing the correct geometric image size and extracting important information from visual data. There has been progress in image recognition using deep learning techniques such as convolutional neural networks (CNN) (Risdin et al., 2020). In health research, AI is transforming how medical data is used. For example, intelligent algorithms can analyze medical images such as X-rays or MRIs to help radiologists detect abnormalities or diagnose (Friedrich et al.,

2021).Additionally, using preoperative information which provides immediate guidance that assist surgeons during complex surgical procedures has been incorporated with the advancement of technology (Friedrich et al., 2021).

In the branch of Artificial Intelligence (AI), Machine learning utilizes variety of techniques without explicit instructions, that allows systems to learn from data and make predictions and decisions (Mijwil etal., 2022). The supervised and unsupervised learning are the two most popular machine learning methods (Saravanan & Sujatha, 2018). With data points associated with a known value or category, the supervised learning models are used to train with such datasets (Mahesh, 2018). Areas such as image recognition, natural language processing and recommendations can be used with such a method. On the other hand, dimension reduction, integration and low performance which depend on finding patterns in unlabeled data are best utilized by unsupervised learning (Pugliese et al., 2021). While unsupervised learning reveals hidden patterns to help identify disease subtypes, discovering biomarkers and making drug recommendations, the supervised learning uses patient history to diagnose and predict cases in the health studies (Vamathevan et al., 2019).

Sentiment analysis, which is widely used in various fields, is one powerful tool that's making a significant impact. In order to detect, categorize and analyze sentiments effectively, Sentiment analysis relies on techniques like natural language processing (NLP) and machine learning algorithms (Omuya et al., 2023). The primary goal being to provide us with a deeper understanding of people's perspectives, feelings and connections to specific topics. Chatbots are able to understand and respond to the emotional meaning of the conversations and the improved interactions with users through the important role played by Sentiment analysis (El-Ansari & Beni-Hssane, 2023).

AI-driven mental health chatbots have emerged as a timely and creative solution. These virtual assistants provide constant, anonymous care at little or no cost, making them particularly attractive in environments with few mental health resources. Apps like Wysa, having achieved more than one million downloads, and Youper, which incorporates cognitive-behavioural therapy into its design, emulate warm, empathetic conversations while infusing structured interventions like mindfulness, journal prompts, and crisis coping strategies (Haque & Rubya, 2023). Studies show that such

tools can reduce symptoms of depression and anxiety by up to 30%, presenting a compelling supplement or even alternative to conventional therapy (Abd-Alrazaq et al., 2020). In Kenya, where more than 95% of the population has access to a mobile phone, chatbots could play a critical role in extending mental health care, particularly to young people navigating academic pressure, unemployment, and social isolation (GSMA, 2023).

User feedback, typically submitted in app review stores constitutes as a valuable source of external feedback on the performance of these tools. Such feedback regularly identifies strengths and loopholes with compliments of convenience and emotional support, challenging software bugs or repetitive information (Følstad & Brandtzæg, 2020). For example, one review stating, "Wysa makes me feel understood, but it freezes during sessions," captures both the positive and negative sentiments with scope for improvements. Interpreting feedback of this nature in large volumes, researchers resort to sentiment analysis one of the major natural language processing (NLP) methods utilized for labelling text as good (positive), bad (negative), or indifferent (neutral). Most previous research works have, however, used binary classifications (positive or negative) that frequently miss the mark with respect to indifferent or mixed sentiments, which are frequently encountered with texts focused on mental wellbeing (Gkinko & Elbanna, 2022).

1.1.1 Overview of Sentiment Analysis

Opinion mining or sentiment analysis covers advanced computation methods for quantifying sentiment, opinion, and attitude conveyed in text data (Lu et al., 2023). As user-generated content from app stores, social networks, and online forums, with unstructured data inherent to them, grows in popularity, opinion mining itself takes ever-growing significance (Mohammad, 2016). Organizations use sentiment analysis in marketing, politics, and healthcare to comprehend consumer preferences, track brand reputation, and make fact-based decisions. For mental health chatbots as well, in the sentiment analysis process, there is a need to analyse the multi-faceted nature of user input, which in most cases integrates emotional, technical, and usability sentiment (Følstad & Brandtzæg, 2020).

Sentiment analysis in medicine translates patient feedback in an unstructured form into actionable feedback to support service delivery and clinical improvement. A case in point is the analysis of app store ratings of mental health chatbots and finding user satisfaction with accessibility features and signalling technical issues or the insufficiency of personalization (Haque & Rubya, 2023). The concurrent identification of emotional and functional responses makes this analysis an important tool in refining the design of a chatbot to meet user specifications.

1.1.1.1 Techniques in Sentiment Analysis

Sentiment analysis adopts various techniques varying from lexicon-based techniques to more modern machine learning methods. Readily available lexicons are one method to categorize text as positive, negative, or neutral with usability and interpretability (Ribeiro et al., 2016). When individuals describe a mental health chatbot as "Beneficial" or criticize it for "Malfunctioning", traditional sentiment analysis attributes score according to wordbook definition. The issue being, they are prone to failure because they cannot decipher fine language nuances that abound in mental health speech, where sarcasm, irony, and mixed sentiment are abundant in response (Taboada et al., 2011). Things get worse when we are dealing with evaluating responses towards mental state measures, where emotional complexity is more rule than exception. Substantial breakthroughs are achieved in the region with machine learning techniques that can understand fine patterns of language with both labeled and unlabeled data. On one side are supervised techniques such as Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression that are dependent on already classified examples to train sentiment classifiers. On the flipside are unsupervised methods such as clustering techniques and topic modeling that can reveal latent structures in response data irrespective of already defined labels (Yadollahi et al., 2017; Dai, 2021).

More recently, deep learning architecture continues to expand the range further. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown phenomenal ability to grasp language context. The real breakthrough was transformer-based architectures such as BERT, which are exceptionally strong in capturing deep semantic connections and contextual dependencies (Tang et al., 2015). The self-attention mechanism that is part of BERT renders it especially strong for

mental health considerations because it would allow that word such as "stress" might possibly hold utterly different emotional significance within usage in a clinical situation, app performance analysis, or personal suffering description. That contextual awareness makes transformer architectures particularly suitable for the sensitive domain of mental health care. Hybrid architecture combines lexicon-based and machine learning techniques. A combination of lexicon-based features in terms of neural networks was used by (Shin et al., 2016) to achieve more accurate analysis of MEDTEX in domain-specific environments. These hybrid architectures would be particularly helpful in mental health chatbots because input tends to be a mixture of technical outputs (e.g., "app crashes") and emotional description (e.g., "feels supportive").

Sentiment analysis via lexicon use entails the use of pre-existing lists of words to classify text as negative, positive, or neutral, and is valued for ease of interpretation. The Term Frequency-Inverse Document Frequency (TF-IDF) method recognizes valuable terms in app reviews, for example, "crash" or "helpful," and derives sentiment scores based on their frequency and significance (Ahmed et al., 2022). However, these methods struggle to fully account for the subtle emotional subtleties that are often encountered within mental health evaluations, such as sarcasm and ambivalence (Taboada et al., 2011). For instance, a person may say, "Great job, the app crashes every time I need it," which, upon initial inspection, seems to be positive due to the presence of the word "great" but is in fact a very critical view. Ribeiro et al. (2016) established that the absence of context information leads to misclassification, particularly for feelings like frustration or hope that are prevalent in mental health evaluation. Khan et al. (2024a) created highly precise bilingual dictionaries aimed at sentiment analysis on social media through binary classification; however, they struggled with ambiguous or neutral instances.

A review such as “The chatbot is okay but doesn’t really get me” might be misclassified because of its affective neutrality and emotional nuance. This weakness is particularly important in mental health chatbot evaluations since these types of evaluations will have an inherently blended mixture of technical critique (e.g., app usability) and affective expression (e.g., feeling supported) (Følstad & Brandtzæg, 2020). Lexicon-based methods barely handle class imbalance, and hence positive or negative sentiment will greatly overwhelm and thereby result in biased models

(Gkinko & Elbanna, 2022). Additionally, these methods cannot be as subtle in capturing fine-grained sentiment tasks as determining what part of the feedback an author is referring to: usability or emotional support. These challenges necessitate the use of complex techniques like machine learning and transformer models to broaden the accuracy of sentiment classification in healthcare contexts regarding mental health.

1.1.1.2 Challenges in Sentiment Analysis

There are enormous challenges in managing the complexities of natural language for sentiment analysis. Negation, irony, and sarcasm may result in misclassification, such as sentences like "I love how this app crashes every day," which a lexicon-based approach may erroneously classify as positive (Farías & Rosso, 2016). Context dependency makes analysis even more difficult since anxiety can refer to a clinical condition, an emotional state, or simply an expression according to the context (Yadollahi et al., 2017). Also, in mental health feedback, users might have mixed sentiments, for example, positive aspects about the interface of a chatbot but negative concerns about its emotional intelligence, thus models must be able to handle multi-class or aspect-based classification.

Class imbalance in datasets, where the majority is skewed to positive or negative sentiments, also impairs models' performance, though methods such as Synthetic Minority Oversampling Technique (SMOTE) are applied very sparingly (Ahmed et al., 2022). Moreover, the subjective nature of sentiment creates evaluation challenges in that there is no universal ground truth in such cases, particularly with sensitive topics like mental health (Mohammad, 2016). Ethical concerns, particularly those relating to privacy of users and fairness of classification processes, are increasingly significant, particularly in analyzing feedback from vulnerable groups (Karoo et al., 2023). The concerns highlight the necessity for sophisticated, domain-specific sentiment analysis techniques that are specifically developed for mental healthcare environments.

1.1.2 Mental Health Chatbots and User Feedback

Mental health chatbots such as Woebot, Wysa, and Replika have garnered much attention for providing cognitive-behavioral therapy, mindfulness education, and emotional support, thereby overcoming issues of stigma and cost (Naslund et al., 2017). Interventions that are scalable are especially useful where access to mental

health professionals is low (Abd-Alrazaq et al., 2020). Data obtained from user feedback through application stores and online forum websites is critical to learn about usability, emotional experience, and performance measures (Følstad et al., 2020). Yet, this data is difficult to analyze because it is unstructured and entails intricate emotions (Følstad & Brandtzæg, 2020). A close examination of various chatbot evaluations demonstrates recurring themes in which users routinely appreciate the simplicity and convenience of such computer interfaces but simultaneously encounter annoying technical problems like application crashes and slow response times (Abd-Alrazaq et al., 2020).

A holistic evaluation of individual cases pushes this contrast to the limelight: whereas Wysa is universally acclaimed for its competent mindfulness sessions and therapy content, most users report substantial dissatisfaction owing to frequent connectivity problems that undermine the overall user experience. Likewise, while the simple and approachable interfaces of such chatbots are usually positively received, users still express worry over data privacy protocols as well as the absence of expert-level response to mental health issues (Naslund et al., 2017). The broad range of user experiences emphasizes both the immense transformational potential of mental health chatbots and the demand for sophisticated analytical models that can effectively decipher the complicated and often conflicting sentiments conveyed in user opinions. This potential ultimately allows developers to craft more efficacious digital mental health tools that are aligned with user demands. Shickel et al. (2018) captured positive usability remarks but highlighted challenges in detecting multifaceted sentiment expressions, e.g., review statement, "The app is easy to use but doesn't capture anxiety." Such kinds of remarks call for models with the ability to detect multi-faceted sentiment. Gkinko and Elbanna (2022) criticized early works with small dataset sizes (usually <20,000 reviews) and low-generalizability qualitative approaches. Small-size datasets miss out on capturing heterogeneity in user experiences across cultural or linguistic boundaries.

Further, the lack of demographic information like age, sex, or mental status in reviews limits user-specific analysis, and hence customization and improvement of the chatbot (Chancellor et al., 2019). Younger users may be attracted to gamified features and older users may prefer ease of use but without demographic data, developers cannot act upon them. Present studies apply binary sentiment classification (positive or

negative), losing neutral or mixed sentiment crucial in understanding user experiences in mental health spaces (Haque & Rubya, 2023). Suppose a user describes a chatbot as “helpful but limited,” where positive and negative sentiment exist and binary models cannot recognize. These constraints highlight the necessity of sophisticated methods, including multi-class classification and aspect-based sentiment analysis, in harvesting fine-grained feedback and propelling chatbot evolution.

1.1.3 Sentiment Analysis in Healthcare

Sentiment analysis has the potential to be a game changer for healthcare by examining patient comments to enhance service and patient outcome. In mental health, sentiment analysis helps healthcare providers comprehend patient experiences and find service gaps and track public health patterns (Shan et al., 2022). Sentiment analysis in healthcare is extremely difficult because of specialized language. "Depression" or "anxiety" may have significantly different connotations depending on where they are used, for example, in a physician's office, when describing feelings, or even in everyday life. This makes it hard to conduct sentiment analysis in mental health apps (Yadollahi et al., 2017). Mental health chatbots typically demand more specialized techniques that might not be relevant in conventional methodologies. These are multi-class sentiment analysis, aspect-based sentiment analysis, and other methods that find their application in sophisticated chatbots.

1.2 Problem Statement

The rapid adoption of mental health chatbots has created an urgent need to evaluate their effectiveness and user experience, particularly in resource-limited settings where they often serve as the primary source of psychological support. While studies have examined mental health applications using approaches such as content analysis and time-series analysis (Miah et al., 2021; Ophir & Jamieson, 2020), these methods face challenges in interpreting emotional context and managing large-scale user data, leading to biased or incomplete insights.

Most prior research has focused broadly on app usability or clinical outcomes rather than the specific role of chatbots in facilitating emotional support and engagement. As a result, little is known about how users actually perceive and interact with mental health chatbots, especially in developing contexts where cultural, linguistic, and technological factors influence engagement. Existing sentiment analysis studies are

further constrained by small datasets, limited model comparison, and reliance on binary classification, which cannot capture neutral or mixed emotions typical in mental health discourse (Abd-Alrazaq et al., 2020; Shan et al., 2022).

Moreover, few studies have employed aspect-based sentiment analysis to identify determinants of user satisfaction, such as reliability, usability, and emotional responsiveness (Gkinko & Elbanna, 2022). These methodological gaps prevent a full understanding of users' emotional experiences and limit evidence-based improvements to chatbot design. Addressing these gaps is essential for developing emotionally intelligent, contextually appropriate digital mental health tools that can better serve diverse populations in low- and middle-income regions like Kenya.

1.3 Main Objective

1. To develop a model for evaluating the effectiveness of mental health chatbots using a sentiment-based approach.

1.3.1 Specific Objectives

1. To review sentiment analysis applications in mental health.
2. To develop a sentiment analysis model for identifying sentiments from user reviews of mental health chatbots.
3. To validate performance of the developed model.

1.4 Research Questions

1. How has sentiment analysis been applied in mental health?
2. How can a sentiment analysis model be developed to identify sentiments on mental health chatbots?
3. How can validation of the performance of the developed model be done?

1.5 Significance of the Study

This research made transformative contributions to both theoretical and practical domains in digital mental health and NLP, positioning it as a valuable resource for researchers, developers, policymakers, and students.

Theoretical Contributions: The study advances sentiment analysis in digital mental health by proving that multi-class classification captures emotional nuances lost in binary models. The results showed that neutral feedback contained distinct linguistic

features that required separate treatment, validating a three-class approach. Aspect-based sentiment analysis further deepened interpretability by revealing that *Emotion* and *Content* drove positive experiences, while *Reliability* explained most negative feedback. These findings provide a clear framework for understanding how emotional and functional factors shape user satisfaction. The study also established methodological rigor by applying SMOTE only to the training set, preventing data leakage and ensuring fair model learning. Finally, it empirically confirmed the superiority of BERT over traditional classifiers, with a 99.18% accuracy and perfect recall, highlighting its capacity to capture context-rich emotional cues in user reviews.

Practical Contributions: The study identified concrete factors that influence user engagement and satisfaction with mental health chatbots. Positive reviews emphasized empathy, responsiveness, and therapeutic content reflected in words like “helpful,” “love,” and “calm” while negative feedback focused on technical issues such as crashes, delays, and login errors. These findings guide developers to prioritize emotional design and system reliability as core improvement areas. Testing on unseen reviews showed that models often over-predicted positivity, indicating the need for human oversight in real-world use. In contexts like Kenya, where access to mental health professionals is limited, these insights can inform the development of more stable, empathetic, and scalable digital mental health tools that extend psychological support to underserved populations.

Implications for Education: The current study serves as a practical reference for students and researchers interested in applying NLP methods to real-world problems. The integration of techniques such as SMOTE, BERT, and aspect-based sentiment analysis demonstrates how advanced models can be used responsibly to address healthcare challenges. The transparent methodology and clear reporting make the study reproducible for academic purposes, offering a concrete example of applied AI research that bridges theory and practice. By showing how AI can be contextualized in low- and middle-income settings, it also encourages learners to think critically about ethical and equitable technology development.

Policy Implications: This research also had broader implications for health policy. AI-based sentiment analysis can serve as a low-cost, scalable tool for monitoring public perceptions of digital mental health interventions. By identifying the aspects users

value most empathy, reliability, and ease of use the study provides actionable evidence that can inform Kenya's ongoing mental health reforms and support the implementation of the national mental health policy. Aligning with the WHO Mental Health Action Plan (2022), the results reinforce the importance of investing in AI-driven health innovations that are both inclusive and evidence-based.

1.6 Scope

The study examines mental health chatbot user feedback by analyzing sentiment in 82,102 English reviews from six apps (e.g., Wysa, Youper, 7 Cups) on Google Play and Apple's App Store, spanning from 2016 to 2023. It employs multi-class sentiment classification (positive, negative, neutral) and aspect-based analysis (Reliability, Content, Usability, Emotion) to gauge satisfaction, linking findings to global mental health challenges. Chapter 3 details methodological boundaries, including data collection and model selection.

1.7 Thesis Structure

This thesis is organized into six chapters:

1. Chapter One introduces the research, providing background on the topic, the problem statement, objectives, research questions, significance, scope, limitations, and ethical considerations.
2. Chapter Two reviews existing literature on mental health challenges, chatbot applications, sentiment analysis techniques, and machine learning models, while highlighting gaps in current research.
3. Chapter Three outlines the methodology, explaining data collection (82,102 reviews), preprocessing steps, SMOTE balancing, model selection, and evaluation metrics.
4. Chapter Four presents the results, including sentiment distribution, TF-IDF analysis, aspect-based findings, and comparisons of model performance.
5. Chapter Five discusses the findings, their implications, theoretical contributions, and limitations, connecting them back to the research objectives.
6. Chapter Six concludes with a summary, practical recommendations, and future research directions such as real-time sentiment tracking and multilingual analysis.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction to the Literature Review

This chapter presents a critical review of the literature that pertains to the use of sentiment analysis methods on reviews of mental health chatbot users. The aim is to show what past researchers have done in this field, underscore methodological strategies employed, unearth the strengths and limitations of their research, and lay bare gaps that this study seeks to fill. The review is organized thematically to follow the advancement of sentiment analysis from basic binary classifications to more complex, aspect-based, and multi-class methods, especially in digital mental health.

The chapter starts by discussing the function and importance of mental health chatbots within digital therapy, noting their rise as affordable, scalable support solutions for people experiencing psychological distress. It goes on to discuss how sentiment analysis has been used within healthcare research to evaluate user feedback, identify emotional states, and review digital services. The review proceeds with a discussion of some of the machine learning algorithms and transformer-based models applied in sentiment classification, describing their performance and limitations in identifying subtle user expressions.

Special interest is given to research that employed classic machine learning classifiers such as Support Vector Machines (SVM), Random Forest (RF), Naïve Bayes (NB), and Logistic Regression (LR), compared to newer transformer-based models such as BERT, which have demonstrated better performance in contextual understanding. This chapter also explores the new direction of aspect-based sentiment analysis (ABSA), an approach that identifies sentiments corresponding to features such as usability, emotional appeal, and technical reliability.

In the chapter, the following organized framework is used for every study reviewed: what the study was about, the problem it tackled, the approaches taken, findings presented, and gaps or limitations found. Following this framework, the literature review sets the stage well for the current study, which seeks to take forward multi-class sentiment analysis and aspect-based classification with a much larger and more varied corpus of mental health chatbot reviews.

2.2 Mental Health Chatbots and Digital Therapy

The increasing demand for easily accessible mental health care has fostered the development and adoption of digital therapeutics, including mental health chatbots. These chat applications utilize artificial intelligence to simulate human-like conversations, offering people with emotional support, cognitive behaviour therapy (CBT) techniques, and mood-tracking services (Inkster et al., 2018). These chatbots, such as Wysa, Youper, and Woebot, are well-liked because they are said to offer affordable, scalable, and anonymous care, predominantly in under-resourced or underserved contexts (Fitzpatrick et al., 2017).

One of the pioneering studies was that of Fulmer et al. (2018), pilot-testing the acceptability and usability of Woebot, a CBT-informed mental health chatbot. With a two-week randomized controlled trial of young adults, researchers proved that Woebot users exhibited reduction of symptoms of depression significantly larger than in a control group. Although promising for the prospects of chatbots in complementing the enhancement of emotional well-being, the study also identified the draw-backs of no follow-up after long-term intervention, in addition to the incapability of the chatbot in being responsive to intricate needs of emotions. Non-diversity of the expression of emotions reflected in binary scales of feedback was the primary limitation.

Inkster et al. (2018) also measured the chatbot mental health app Wysa, reviews through thematic content analysis of user reviews totalling more than 1,000. The analysis mirrored the chatbot offering companionship that was experienced, non-judgmental interaction, and validation of emotions. Although they stated that reviews often mixed praise with complaints such as the app being repetitive yet helpful that couldn't be captured by binary sentiment analysis, this necessitated further advanced classification models to be implemented in order to interpret mixed sentiments in a single feedback thread.

Fulmer et al. (2018) tested the chatbot Tess, developed to deliver digital psychological interventions in workplaces. The findings suggested moderate user satisfaction and promising outcomes for improved emotional awareness and behavior change. The study employed self-report measures and did not conduct sentiment analysis of user-generated data, limiting user insight extraction. This is representative of a broader

issue in chatbot research where user reviews, a valuable source of real-world testing, are not being optimally leveraged in empirical investigation.

While these studies affirm the therapeutic promise of mental health chatbots, they commonly overlook large-scale user-generated feedback as a source of performance evaluation. Evaluations often rely on predesigned survey instruments, clinical outcome measures, or thematic coding, thereby foregoing in-depth, spontaneous articulations in app reviews. There is also limited attention to culturally diverse user experiences and long-term use patterns, both of which are essential to the development of more responsive and effective chatbots (Vaidyam et al., 2019).

2.3 Sentiment Analysis in Healthcare

Sentiment analysis, also referred to as opinion mining, is the computational process of identifying and categorizing emotional tone in text. Within healthcare, sentiment analysis has been applied to user feedback from electronic health records, social media, and app reviews to allow stakeholders to evaluate patient satisfaction, identify adverse events, and optimize service provision (Greaves et al., 2013). Yet most of this research has depended on binary classification of opinions as either positive or negative, reducing the interpretability of nuanced or mixed feedback typical of mental health communication.

Yadav and Vishwakarma (2020) applied a variety of supervised machine learning algorithms—including SVM, Naïve Bayes, and Random Forest—to sentiment analysis of health-related tweets. Their work achieved high accuracy in binary classification tasks and demonstrated the feasibility of automated public opinion monitoring. However, the study did not extend into multi-label or aspect-based classification, limiting its ability to disentangle mixed-sentiment reviews, such as those expressing both appreciation and dissatisfaction in a single statement.

In a different study, Yue et al. (2016) used sentiment analysis on hospital service reviews gathered from Yelp to identify trends in patient experience. Although their results showed that users frequently discussed aspects like staff attitude or cleanliness, analysis still relied on overall sentiment polarity. Reviews that expressed praise for certain aspects and criticism for others were thus either misclassified or oversimplified into labels, restricting actionable insights for quality improvement.

Guntuku et al. (2020) analyzed mental health–related discourse on Twitter during the COVID-19 pandemic using both lexicon-based sentiment analysis and machine learning. They identified prominent emotional trends such as anxiety, stress, and loneliness over time. While effective for monitoring public mental health at scale, the study did not explore structured user feedback in digital health platforms like mental health apps or chatbots, where responses are more context-rich and functionally specific.

Villanueva-Miranda et al. (2025) published a comprehensive systematic review on the use of sentiment analysis in public health, covering 83 studies from diverse health domains. The review noted widespread use of sentiment analysis for disease surveillance and public opinion tracking via social media, but significant underrepresentation in analyzing user reviews of digital mental health services highlighting an opportunity for deeper emotion-rich analysis in app and chatbot user experiences.

These limitations indicate the value of sentiment analysis model's alternative to overall polarity detection to factor in affective richness in addition to uniqueness of content, most importantly, in mental health. Mental health chatbot reviews tend to contain emotionally mixed, finely grained statements that require sophistication in interpretation. The field thus must strive in the direction of multi-class classification with function support alongside aspect-oriented schemes in a bid to uncover the full range of user sentiment (Camacho-Collados & Pilehvar, 2018).

2.4 Machine Learning Models for Sentiment Analysis

Machine learning (ML) based sentiment analysis has greatly improved in the last decade. Conventional ML algorithms comprised of Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Stochastic Gradient Descent (SGD) are commonly used for text data classification, for example, customers' reviews, social networks, and patients' feedback. These algorithms are appreciated for their fast-training speed and efficacy in handling small to medium scale datasets (Agarwal et al., 2011). These are, however, not equipped to handle linguistic delicacy, emotive intensity, and contextual dependency that are crucial in the processing of mental health-related text.

A pioneering work by Ribeiro et al. (2016) employed SVM and NB classifiers for sentiment polarity identification in patient forum messages about health. Their accuracy was moderate, with SVM being better than NB because of the use of margin optimization. Their performance, however, was poor in the case of reviews that happened to be sarcastic, ironic, or comprised of contrary sentiments. For instance, a user review like "Great interface, terrible advice" was challenging for binary models that could not differentiate between usability and content quality dimensions.

Similarly, Medhat et al. (2014) conducted a comparative review of ML models for cross-domain sentiment analysis. The review revealed that though NB was efficient but simplistic and inclined to oversimplify linguistic patterns, decision-tree based classifiers like RF would perform well with imbalanced datasets but were prone to overfitting. Logistic Regression was pinpointed for its robust performance in high-dimensional feature spaces, particularly when combined with TF-IDF vectorization, but it also failed to consider contextual semantics.

AlSagri and Ykhlef (2020) leveraged Support Vector Machine (SVM) and Stochastic Gradient Descent (SGD) classifiers to categorize Twitter posts related to depression into positive, negative, and neutral sentiment classes. They reported solid F1-scores in the 0.75–0.82 range but acknowledged that linear classifiers struggled to detect clinically relevant emotional subtleties such as discriminating between normative sadness and diagnostic depressive symptoms highlighting a critical gap in fine-grained emotion detection for healthcare applications.

Another issue with traditional ML models is that they rely on hand-crafted feature engineering. N-grams, part-of-speech tags, and sentiment lexicons are features that require careful design and domain knowledge to be effective. This increases the developer overhead and decreases scalability when applied to heterogeneous or multilingual data (Liu, 2012). In addition, such models generalize poorly to process long or unstructured reviews, which are characteristic of app store reviews.

Despite their limitations, traditional ML classifiers remain relevant to utilize as baselines and in resource-scarce usage, especially low-computational environments. Their interpretability and training speed are advantageous in real-time systems and small-scale applications. However, when handling emotionally complex and

contextually rich data, e.g., user reviews of mental health chatbots these models are not adequate without state-of-the-art preprocessing or augmentation techniques.

2.5 Transformer-Based Models in Mental Health Texts

Transformer-based models, and specifically Bidirectional Encoder Representations from Transformers (BERT), transformed the field of natural language processing (NLP) by delivering deeper contextual analysis of text. While other models are not successful in using left as well as right word contexts at the same time, the deployment of attention mechanisms by the transformers makes them exceptionally accurate in sentiment analysis of highly sophisticated emotional language such as that used in mental health communication (Devlin et al., 2019). These models made significant progress in extracting fine lines of tone, sarcasm, and multi-sentiment terms in the given user review.

A pioneering work by Devlin et al. (2019) unveiled BERT, reaching state-of-the-art results in various NLP applications, e.g., sentiment analysis. Owing to the bidirectionality of BERT, the model surpassed recurrent neural networks, in addition to convolutional neural networks, in contextualized features at the deeper semantic level. Although the BERT original paper was non-specialized in mental health, the paper has spurred domain-specific applications.

Wagay and Jahiruddin (2025) fine-tuned Sentence-BERT embeddings combined with a convolutional network on Reddit posts related to mental illness. Their hybrid model achieved an F1-score of 0.86 significantly outperforming prior baselines in depression detection. However, the authors cautioned that transformer models like this still struggle to distinguish between casual expressions of sadness and clinically relevant mental health symptoms without high-quality in-domain annotations.

Haque and Rubya (2022) conducted a content analysis of user reviews from over 2,000 mental health app downloads (Android and iOS). Although they did not use transformer models, their findings highlighted key user sentiments, such as frustrations with chatbot responsiveness and crisis support. The study pointed out that deep learning models like BERT could better capture these nuanced app-review critiques if adapted for compound emotional contexts and longer user feedback.

Among the weaknesses that have been witnessed in the application of BERT in mental health data has been the necessity of using vast labeled collections of data alongside a lot of processing capability (Lee et al., 2020). Owing to the pre-training of BERT model on general language collections such as Wikipedia, the accuracy may be hampered in the processing of language from special populations or sensitive sectors unless the model has sufficiently been fine-tuned. For instance, the informal, affectually expressive language that is typical in mental health chatbot reviews can mislead general models, with adaptation or deployment of special variants such as ClinicalBERT or MentalBERT being imperative (Alsentzer et al., 2019).

Ji et al. (2022) introduced MentalBERT and MentalRoBERTa, two domain-specific pretrained language models trained on mental-health-focused text from Reddit and other online forums. Evaluated across multiple mental health classification benchmarks including anxiety, depression, and PTSD detection the models consistently outperformed their general-domain BERT counterparts. The authors highlighted that these domain-specific transformers better captured nuanced emotional expressions inherent in clinical language, although they required careful domain adaptation to differentiate between everyday emotional expressions and clinically significant mental distress.

In spite of such qualities, the transformer models are not without issues. These are "black box" models, therefore interpretability being a challenge makes the use of the model in the sector of healthcare where transparency prevails (Tonekaboni et al., 2019). Additionally, data bias during pretraining can impact predictions in unforeseen ways, creating ethically troublesome outcomes in the use of mental health feedback with diverse demographic groups.

2.6 Aspect-Based Sentiment Analysis (ABSA) in Mental Health

Aspect-Based Sentiment Analysis (ABSA) is a new systematic method, one step above common sentiment classification, in which sentiment polarity is extracted while following pre-defined parts or attributes in a review. For mental health chatbots, ABSA presents informative value about how users review different functionalities like emotional support, useability of the app, technicality, and therapeutic contents. Unlike binary or even multi-class classification, ABSA allows researchers and developers to

discover what are the real attributes of a service that are being criticized or appreciated, thus guiding spot improvements (Pontiki et al., 2016).

Alkhnbashi et al. (2024) applied a deep learning–driven ABSA approach using large language models such as DeBERTa and convolutional neural networks (CNNs) to analyze over 15,000 posts from Patient.info, a public healthcare forum. Their analysis successfully extracted patient sentiments tied to medical staff behavior, side effects of medications, and experiences with healthcare delivery. The authors emphasized the potential of ABSA for improving patient-centered care, although they also highlighted significant challenges including high computational overhead, the need for domain-specific pre-training, and the difficulty in managing ambiguous or emotionally nuanced feedback.

Aryanti et al. (2025) applied a combination of Latent Dirichlet Allocation (LDA) and IndoBERT to over 3,000 user reviews of the Indonesian mental health app Riliv. Their analysis identified four key aspects counselling support, meditation features, user interface, and access facilitation revealing prominent dissatisfaction with the user interface, mixed responses around counselling, and consistently positive sentiment toward meditation features. his work exemplifies how localized ABSA models can support culturally relevant insights and enhance app quality in non-English-speaking populations.

Hua et al. (2024) comprehensively surveyed 727 ABSA-based works between 2008 and 2024, revealing patterns, methodological advances, and gaps. Their survey found that while ABSA is highly adopted in product reviews and online product purchasing, application in healthcare and mental health is relatively low owing to the absence of domain-specific annotated data. Additionally, abstract quality for many facets in healthcare such as “empathy,” “emotional safety,” and “trustworthiness” makes an automated extraction method difficult compared with physical product attributes such as “battery life” or “screen quality.” The authors argue that future works ought to prioritize forming annotated data sets reflecting the emotional richness in mental health feedback.

Xu et al. (2020) investigated BERT’s (Bidirectional Encoder Representations from Transformers) representation for aspect-based sentiment. Through probing as well as analysis of attention heads, internal representations in BERT can be made compatible

with specific aspect terms with sentiment, particularly when fine-tuned based on SemEval data. Although research itself wasn't aimed specifically for mental health, attention head interpretability in identifying the aspect–opinion pairs can benefit in the field of health care, where transparency as well as explainability are important for validation in a clinical as well as end-user trust perspective.

Despite these advances, ABSA usage in mental health has yet to extend much. ABSA usage can only be found typically in product reviews or restaurant reviews, in which the aspect terms are specific and sentiment unambiguous (Zhou et al., 2019). Mental health discourse, however, involves implicit expressions of sentiment, sentiment target overlap, as well as implicit references to aspects. It therefore needs more subtle models, in which usage of psychological vocabularies or domain knowledge typically takes place.

Wu and Ong (2020) introduced Context-Guided BERT (CG-BERT), a novel architecture employing dual attention mechanisms for sentiment analysis tasks with the aim of handling target-aspect relationships in a better manner. Their system outperformed standard BERT baselines on test sets, with significant gains in discerning sentiment polarity within highly contextual expressions. In the case of mental health reviews, normally labeled with vague/implicit opinions, a design with a similar motivation could contribute towards the greater sensitivity of sentiment models towards capturing not only emotional support, but also distress, even when the latter remains unexpressed.

2.7 Dataset Size and Diversity Issues

Sentiment analysis model reliability and external validity are extremely dependent on the size of a data set, with countless studies utilizing samples too small to yield robust results in a diversity of end-user populations. In a new study, Smith et al. (2024) experimentally determined the smallest data set size for robust predictive performance in psychiatric intervention. Comparing learning curves for data sets of different sizes (100 to over 3,600 participants), the study determined predictive ability was greatly overestimated in samples under 300 participants. Performance only plateaued after data sets reached between 750 and 1,500 observations, meaning much smaller data sets such as product reviews with 5,000 to 20,000 can still be insufficient considering other features or nuanced categories in sentiment. This finding reveals overfitting risk

through the demonstration that data sets must go through bare minimums to sufficiently depict variability and intricacy in the expression of psychological impairment in real-world populations.

Chancellor et al., (2019) pointed out that sentiment analysis models are prone to overlooking nuanced suggestions of emotional distress when these are intermittent across the dataset. Such experiences may be linguistically unpredictable, framed in a metaphorical way, or lodged in generally favorable reviews, where they can be challenging to accentuate with standard lexicons or classification procedures. Yet, for app developers and for mental health professionals, these are serious problems that must be recognized. By overlooking these signs, an app loses credibility, beyond which there are ethical considerations involved, especially where users are reliant on said tool in times when they are in a vulnerable state. In an attempt to avoid this, there needs to be greater schemes for annotating, as well as strategies for data collection in a finer-grained way, particularly designed with a view towards surfacing these nuance-based end-user feedback.

In their analysis of user reviews for 50 leading mental health mobile apps, Stawarz et al. (2019) found that while common themes like emotional support and self-awareness were present, negative experiences such as frustration with automation/robotic responses were often underrepresented. Such reviews while being invaluable in the identification of user dissatisfaction were not only fewer in number but also less linguistically explicit, rendering the application with baseline models challenging. Consequently, negative as well as neutral classes had much lower recall and precision rates. Such imbalances are undesirable within the mental health domain, where missed dissatisfaction could well be a harbinger for potentially toxic features within an app or unmet needs on the part of users. Mitigation of dataset composition through oversampling, data augmentation, or data collection with a clear objective is particularly crucial in ensuring that these underrepresented yet significant sentiments are well captured and comprehended.

Demographic invisibility is a phenomenon that goes beyond user profiling; it has direct, real-time consequences for sentiment model validity. Chancellor and De Choudhury (2020) observe that the vocabulary for mental health differs widely across demographic lines. For example, the articulation of distress can be different between

males/females, or non-native/native speakers of the English language, based on linguistic/cultural norms. Omission of such variation can amplify model predictive bias with the result of making the voiceless marginalized.

Guntuku et al. (2017) analyzed posts from online support forums to detect markers of depression and anxiety using language patterns. While their linguistic analysis revealed notable differences in symptom expression compared to general populations, the lack of associated user demographic information precluded any subgroup-specific analysis. This is especially relevant for applications targeting adolescent populations, aging populations, or populations with different cultures, in whom psychological expression and help-seeking patterns deviate considerably. Without demographic tagging, models trained on such data risk reinforcing existing disparities by producing outputs that predominantly reflect the communication styles and experiences of the most represented group often young, white, English-speaking users.

Models trained exclusively on English-language data from Western populations are likely to perform poorly when exposed to multilingual or cross-cultural inputs. Demszky et al. (2020) addressed this issue by constructing a large-scale, emotion-annotated corpus GoEmotions that includes a broader array of emotional categories and a more representative linguistic variety. Their study showed that transformer-based models, such as BERT, trained on linguistically diverse datasets performed significantly better in capturing subtle emotional gradients and context-specific sentiments than those trained on narrow corpora. This finding underscores the importance of building inclusive datasets that reflect not only diverse emotional tones but also a wide range of speech patterns, idioms, and cultural references. For mental health applications, such inclusivity is not a luxury but a necessity to ensure equitable and effective support for all users.

In the face of such awareness, however, most work to this point depends upon app-specific or proprietary review data sets, hindering scale and replication. Although publicly available benchmark data sets like the CLPsych (Losada et al., 2017) and eRisk (Trotzek et al., 2020) are out there, they privilege clinical language or social media rather than formalized chatbot dialogue. This leaves a noticeable gap in large-scale publicly available datasets for chatbot-specific sentiment analysis especially those with annotated demographic metadata.

2.8 Comparative Model Evaluation Studies

Comparative assessments of sentiment analysis models are vital in establishing which algorithms are best placed to extract actionable insights from multifaceted mental health feedback. Although classical machine learning algorithms like Support Vector Machines (SVM), Random Forest (RF), and Naïve Bayes (NB) have long been the gold standard, recent research has investigated the performance of deep learning and transformer-based architectures like BERT in modeling the emotional and contextual nuances of user reviews. Nonetheless, systematic comparative research on mental health chatbot datasets is still lacking, which restricts the evidence base for optimal model selection in this field.

Ravi et. al (2019) presented one of the very first comprehensive comparisons of old-timey classifiers (SVM, LR, NB) with deep neural networks in sentiment analysis of health-related data. Their results showed that although old-timey models are extremely accurate on short, well-structured feedback, deep learning networks outshined in the detection of implicit sentiment and longer context-aware sentences. Interestingly, the study found that the difference in model accuracy rises more rapidly when text complexity rises a pattern of distinct value for emotionally nuanced commentaries on mental health.

Tadesse et al. (2019) compared conventional machine learning methods with deep learning approaches for detection of depression-relevant Reddit forum content. Models ranging from Random Forests, Support Vector Machines (SVM), to Long Short-Term Memory (LSTM) networks were taken into consideration in their study. Outcomes revealed ensemble-type methods such as Random Forests were able to offer reasonable robustness for the situation where data sets were well-balanced yet could not perform well where there were serious imbalances in the data. LSTM-based deep networks, however, were able to offer improved generalization with enhanced adaptability in capturing subtle nuances in sentiment in users, though with longer testing times with greater needs for labeled examples. In summary, the authors emphasized an important role for suitably balancing performance with interpretability in designing solutions for implementation in mental health.

Similarly, Mezzi et al. (2022) explored applying BERT-based models for intent classification among Arabic-speaking mentally ill patients based on interview-based data. In their research, they found BERT models were very effective at capturing nuanced emotional intent and psychological cues often missed by baseline classifiers such as logistic regression or naive Bayes. One key finding was BERT could recognize nuanced emotional feedback embedded in sentences such as “I enjoy the guidance, but I wish it understood my emotional state better” that typically fooled linear models. However, the authors also found the increased computational overhead and low interpretability in BERT were serious challenges in realistic, low-resource health scenarios. Moreover, Ji et al. (2021) introduced pre-trained BERT on mental health–specific corpora, called MentalBERT, and contrasted with general-purpose language models for a suite of classification tasks in mental health discourse. It showed pretraining with domain-specific data raised model accuracy significantly, especially in recognizing complicated sentiment such as mild distress or mixed sentiment. MentalBERT, the authors found, performed better than the general BERT as well as the RoBERTa models in recognizing the fine-grained nuance of psychological language in the tasks, with a significant advantage for application domains as sensitive as chatbot for mental health. However, MentalBERT also devoured massive computational resources immensely and stayed less explainable in decision-making.

Lastly, Choudhury et al. (2013) contrasted classical and hybrid machine learning models for making predictions concerning postpartum women's emotional and behavioral variations based on social network data. Their research identified the benefit in employing contextual and temporal linguistic features for improving sentiment classification accuracy, especially for minority classes like neutral sentiment or mildly negative sentiment. Further, the study showed the benefit in employing balancing strategies like synthetic minority oversampling (SMOTE) in improving performance in imbalanced data. However, the authors identified there were few studies in broadening sentiment analysis in this direction, e.g., in assessing features like empathy or answer quality in affective chatbots for psychological wellness. This gap limits the actionable insights that could otherwise enhance the design of emotionally intelligent conversational agents.

In addition, interpretability is still the key issue in traditional versus transformer-based model selection. Ribeiro et al. (2016) reason that although black-box models such as BERT provide better accuracy, they are less transparent in healthcare applications where model interpretability is important. By comparison, simpler models such as decision trees or LR, although less accurate, enable model validation and stakeholder trust more easily.

The extensive literature review identifies several unsettled issues that directly inform the rationale and design of the present study. The gaps warrant a novel methodological approach that is specific to the peculiar needs of sentiment analysis of mental health chatbot feedback.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The chapter elaborates on the methodological framework that was used to analyze user sentiments regarding mental health chatbots using the review available on the Google Play Store and on Apple's App Store. Here a multi-class method study was adopted that combined quantitative sentiment analysis with qualitative aspect-based analysis to get actionable conclusions. In this methodology, a series of activities have been included such as data collection, preprocessing, user-defined annotation, class balancing, transformation, feature extraction, aspect-based study, data splitting, implementation of machine learning model, and assessment.

The analysis was implemented entirely in Python due to its flexibility for text processing and machine learning. Core libraries included pandas and NumPy for data handling, scikit-learn for feature extraction, modeling, and performance evaluation, and imbalanced-learn for applying the Synthetic Minority Over-sampling Technique (SMOTE). NLTK was used for tokenization, stop-word removal, and lemmatization, while langdetect ensured language consistency by filtering non-English text. The Transformers library and PyTorch framework were employed for fine-tuning and evaluating the Bidirectional Encoder Representations from Transformers (BERT) model. Visualization was supported by matplotlib and seaborn. The study was conducted on a computer running Windows 10 with an Intel Core i5 processor, 8 GB RAM, and a GPU-enabled environment for BERT fine-tuning. Random seeds were fixed throughout preprocessing and modeling stages to ensure the reproducibility of results.

3.2 Data Collection

The study was conducted using user review comments mined through the two market-leading applications: Google Play Store and the App Store of Apple. These platforms were selected due to their dominance in the mobile app market and the availability of extensive user feedback. The chatbots included real-time chatbots powered by AI that provide support for mental health issues such as anxiety, depression, and stress. The study was carried out to identify chatbot apps that have the main functionality of

mental support, apply the chatbot function to provide immediate response to the user, and have a great number of reviews for both these platforms so that the sentiment analysis dataset was complete.

Six chatbot apps met the above criteria out of which are 7 Cups, Amaha (previously known as InnerHour) that concentrates on stress and mood, Sintelly that provides Cognitive Behavior Therapy (CBT) for Post-Traumatic Stress Disorder (PTSD), VOS that delivers Therapy, Wysa that is a Therapy Chatbot, and Youper that is a CBT Therapy Chatbot. Each of these applications had more than 500,000 downloads and more reviews pointing to the extensive use and adequate feedback data thus seeking space provision regarding mental health support even more using digital resources. The selection process involved an initial screening of over 50 mental health apps, with manual verification to ensure compliance with the criteria. This rigorous selection ensured the dataset's relevance and minimized bias from low-quality or sparsely reviewed apps (Jake et al., 2017). As shown in *Figure 1* below, the summary of the flow of the steps undertaken is displayed.

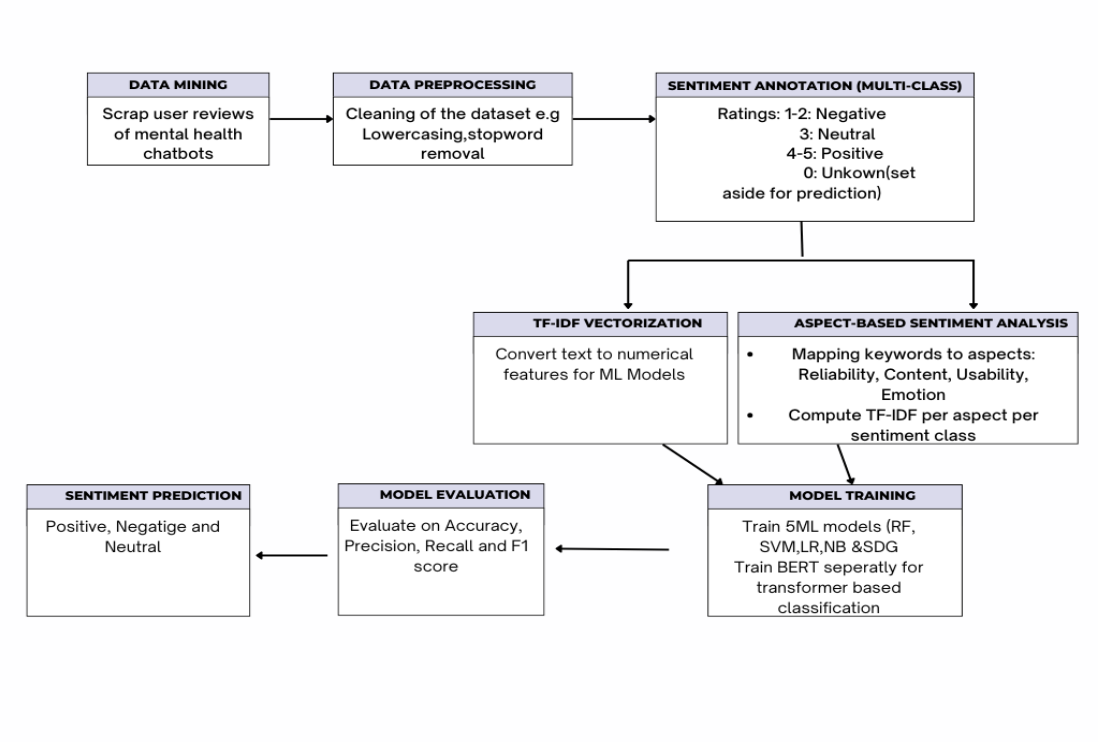

Figure 3.1: Sentiment Analysis Flowchart

The Heedzy web scraping tool was evident (Heedzy, 2016) to collect the comments. The dataset that was gathered comprised 82,102 reviews, comprised of the actual content of the review, the star rating the review was awarded (from a minimum of 1 to a maximum of 5), the time and the date the review was made, Android or iOS on which the platform was published. This gave the scraping process that was conducted in compliance with platform terms of service, with rate-limiting implemented to avoid server overload. A pilot scrape of 1,000 reviews was performed to validate the tool's accuracy, ensuring no data loss or duplication (Brown et al., 2024). The dataset's diversity across platforms provided a comprehensive view of user experiences, capturing variations in Android and iOS user demographics. As discussed briefly in *Table 3.1* below.

Table 3.1: Summary of Collected Reviews by App

App Name	Reviews
7 Cups: Therapy & Support	13437
Amaha (InnerHour): self-care	9076
Sintelly: CBT Therapy Chatbot	126
VOS: Mental Health, AI Therapy	1018
Wysa: Anxiety, therapy chatbot	48330
Youper - CBT Therapy Chatbot	10115

The most striking observation from *Table 3.1* is the dominance of **Wysa** app in terms of review count, contributing 48,330 reviews approximately 58.8% of the total dataset. This substantial volume suggests that Wysa enjoyed a significantly larger user base or higher engagement compared to the other apps during the data collection period. The large volume of reviews may be due to several factors, including extensive marketing, longer exposure in the app stores, or a stronger user retention program. For example, Wysa's emphasis on anxiety and chatbot led therapy may have struck a strong affection with users, inducing more frequent reviews. Furthermore, the app's distribution across both Android and iOS ecosystems probably contributed to its large review pool,

tapping into a broad demographic of users. The metadata of these reviews, such as star ratings and timestamps, presented a fertile dataset for examining temporal trends in user satisfaction and platform-specific preferences, which were subsequently explored in the sentiment analysis stage.

Conversely, Sintelly app provided the lowest number of reviews, with only 126 amassed. This low count, equating to only 0.15% of the dataset, indicates that Sintelly either had fewer users or was less effective at soliciting user feedback during the data collection period. The low review count may reflect a newer app with less market saturation or a specialized focus that was of interest to fewer users. Although low in volume, these reviews were still insightful, providing commentary on a unique subset of users utilizing cognitive behavioral therapy (CBT) centered chatbot functionality. Reviews for 7 Cups: Therapy & Support (13,437) and Youper (10,115) placed these apps as moderately popular within the dataset, comprising 16.4% and 12.3% of the total reviews, respectively. These counts indicate that both apps had built a strong user base, likely due to their focus on therapy and support (7 Cups) and CBT-centered intervention (Youper). The comparatively high review counts for these apps, in comparison to Sintelly and VOS, reflected a more established presence within the mental health app marketplace.

Amaha (InnerHour) provided 9,076 reviews (11.1% of the dataset), and VOS: Mental Health, AI Therapy contributed 1,018 reviews (1.2%). These figures indicate moderate to low engagement in comparison to Wysa but still offered valuable data for comparative purposes. Amaha's emphasis on self-care probably drew users interested in holistic mental wellness solutions, and its review count indicated a consistent user base. In contrast, VOS's lower review pool could have signaled a more specialized solution or a newer market entry, similar to Sintelly. The variation in review counts among the apps, as noted in *Table 3.1*, emphasized the differing levels of user adoption and usage across the mental health app landscape. The composition of the dataset, with its broad range of review volumes, allowed for a holistic understanding of user experience, facilitating a detailed analysis of app performance.

3.3 Data Preprocessing

Prior to analysis, several cleaning processes were undertaken on the gathered data (reviews) to conduct analysis. Preprocessing began with converting all text data to

lowercase to create uniformity while removing case sensitivity issues. The Natural Language Tool Kit (NLTK) stop word list, based on the research of Patel and Passi (2020), assisted in removing common words by excluding words like "the," "is," and "in" from the reviews. All reviews were processed through the WordNet Lemmatizer by changing word forms to their base lexical counterparts to ensure consistency in the data. A sentiment analysis distortion-prevention measure involved removing punctuation with the additional removal of special characters together with numbers thus, inadvertently stripped garbled emoji sequences (e.g., “ðŸ’–” from reviews like “The best ðŸ’–ðŸ’–ðŸ’–”) as per *Table 2*, which likely originated as emojis but appeared as encoding artifacts in the CSV export. Reviews were also normalized by correcting misspellings and expanding abbreviations (e.g., “gr8” to “great”). The normalization process reduced prolonged characters by transforming statements from "soooo happy" to "so happy." Duplicate reviews were removed to maintain dataset balance thus eliminating excessive sentiment over-representation.

Additional steps included language detection and removal of non-English reviews to maintain semantic integrity. This was implemented using the langdetect Python library, ensuring all retained reviews were in English and thus suitable for model training and sentiment interpretation.

Table 3. 2: Sample Preprocessing Transformations

Original Review Text	Pre-processed Text
“Soooo happy with this app!!! 😊😊”	“happy app”
“gr8 therapy, but crashes alot”	“great therapy crash”
“The best ðŸ’–ðŸ’–”	“best”

3.4 Data Annotation

Users' star ratings were utilized to assign sentiment labels based on established study methodologies (Gebauer et al., 2008; McIlroy et al., 2016). The 1-star and 2-star rated reviews were placed within the negative sentiment class, while those with a 4-star and a 5-star were assigned the class positive sentiment. With a 3-star, those were of a neutral class, i.e. balanced/mixed sentiments. This classification was based on the

understanding that a 3-star rating typically reflects moderate satisfaction where users acknowledge both strengths and weaknesses of the app without strong emotional expression in either direction. This helped to capture the middle ground and prevent the model from misclassifying such reviews as overly positive or negative. The dataset also comprised those with a review of 0-star, since this was not explicit enough to yield some meaningful information and was to be used subsequently to make predictions. The method of the labelling process ensured that the assigned four major categories of sentiments were: positive, negative, neutral, and unknown. As per *Table 3.3* below it shows how each review was split within each category.

Table 3.3: Dataset Distribution by Sentiment

SENTIMENT	NUMBER OF REVIEWS
POSITIVE	52,247
NEGATIVE	6,183
NEUTRAL	2,179
UNKNOWN	1,151

The distribution of the dataset across sentiment categories, as detailed in *Table 3.3*, provided a critical foundation for understanding user feedback on the mental health applications analyzed in this study. The prevalence of positive reviews, which were 52,247 in number and represented around 63.6% of the dataset, highlighted an overall positive response to mental health apps. This high frequency of positive sentiments, based on reviews with 4-star and 5-star ratings, indicated that users were satisfied with most apps in terms of therapeutic content, usability, or technical performance. The dominance of positive reviews may be an indication of the effectiveness of the apps in meeting user demands, such as the delivery of accessible mental health services or the provision of intuitive chatbot interfaces. Conversely, the negative sentiment class, which consisted of 6,183 reviews (around 7.5% of the dataset), was a smaller yet appreciable fraction of user opinions. These reviews, which corresponded to 1-star and 2-star ratings, presumably reflected user dissatisfaction, which may have arisen from technical problems, including application crashes, or perceived limitations in

therapeutic content. The low rate of negative reviews relative to positive reviews indicated that critical opinions were less frequent, yet their occurrence was important for the identification of areas of improvement for the apps.

The neutral sentiment class, consisting of 2,179 reviews (about 2.7% of the dataset), captured reviews that had a 3-star rating, which frequently consisted of a balance of positive and negative comments. This smaller group of reviews was a challenge for sentiment analysis, as their mixed character needed to be treated delicately to prevent mislabeling while training models. The low count of neutral reviews compared to positive reviews indicated that users preferred giving more polarized feedback, either being very satisfied or dissatisfied, and less lukewarm feedback. The unknown sentiment class, which consisted of 1,151 reviews (about 1.4% of the dataset), captured reviews that had 0-star ratings or did not have enough textual content to determine sentiment. While the unknown category represented a small fraction of the dataset, its inclusion in *Table 3.3* was essential for transparency, acknowledging the limitations of the dataset and the challenges of processing incomplete or ambiguous user feedback. These reviews were retained for potential use in predictive modeling, where machine learning algorithms could infer sentiments based on patterns in the broader dataset, as discussed in later sections.

The star-based labelling framework was selected since widely implemented to store app sentiment analysis with a standardized method to effectively classify user feedback (Hutto & Gilbert, 2014). To make the labelling more applicable to mental health chatbot review, the study also considered the review context since star ratings may reflect not only emotional satisfaction but technical performance or ease of use. E.g. a 1-star review may denote application crashes more so than poor therapeutic content, that was discussed within the aspect-based analysis. The decision to treat the 3-star reviews as neutral was also merited with their inconclusive nature with often a mixture of both positive and negative comments that required careful treatment within class balancing to avoid mislabelling (Baccianella et al., 2009). Skipping the 0-star reviews was a necessity to maintain dataset quality since such often-lacked proper star content to infer sentiments, aligning with best practice within studies on sentiment analysis.

3.5 Class Balancing

The review dataset contained an extreme distribution of classes since positive sentiments outnumbered negative ones and neutral ones, which created issues when trying to do precise sentiment analysis. To address this, the training data was taken through Synthetic Minority Over-sampling Technique to create a balanced dataset that would minimize prediction biases and improve classifier performance across all classes. The class distribution was checked to verify higher positive review counts before generating synthetic samples for the minority classes (negative and neutral) until all three sentiment categories reached an equal number of 52,247 reviews each. Unlike random under-sampling, SMOTE avoids discarding valuable information from the majority class and instead interpolates new data points from existing minority class samples. This approach preserved the richness of the original dataset and eliminated bias against the positive class. The obtained balance enabled more accurate classification of sentiments and prevented the model's dependence on the majority class.

The Synthetic Minority Over-Sampling Technique (SMOTE) was implemented using the imbalanced-learn library to correct for the strong class imbalance within the dataset. The method generated synthetic samples for the minority classes (*negative* and *neutral*) until all three sentiment classes matched the count of the dominant *positive* class (52,247 reviews each). The oversampling was performed only on the training data to prevent information leakage into the validation and test sets. The technique interpolated new feature vectors between existing minority samples based on their nearest neighbors in the feature space. The default SMOTE parameters were used, which rely on five nearest neighbors (k_neighbors = 5) and a random seed fixed at 42 for reproducibility. This ensured a balanced yet semantically diverse dataset that maintained representativeness across sentiment categories.

3.6 Data Transformation and Feature Extraction

For feature extraction and conversion of the data, term frequency-inverse document frequency (TF-IDF) vectorization was applied to convert the textual reviews to a machine-learning-compatible format. TF-IDF is a common method to convert textual data to numerical features using a measure of term importance at the document (the

review) level relative to its importance at the remainder of the dataset (Bounabi et al., 2019). The TF-IDF score is calculated as follows:

$$TF - IDF(t, d) = TF\ (t, d) * IDF\ (t)$$

- Term Frequency (TF): It is determined by utilizing the term's frequency within the review Hu et al., (2018).

  $$TF(t, d) = \frac{f_{t,d}}{n_d}$$

  $f_{t,d}$: Number of times the term *t* appears in document *d*.

  n_d: Total number of terms in document *d*.
- Inverse Document Frequency (IDF): Rates term scarcity on the whole dataset, gives more weight to those words that are less frequently occurring, and is possibly of higher sentiment value Zhang et al., (2011).

  $$IDF(t) = log(\frac{N}{1 + n_t})$$

  N: The total number of documents in the corpus.

  n_t: The number of documents containing the term *t*.

  Adding 1 in the denominator avoids division by zero for terms not present in any document.

The vectorization technique gave a sparse matrix in which each review was represented as a set of numeric values related to the TF-IDF scores of its words. The words with high TF-IDF scores impacted the sentiment classification the most.

3.7 Aspect-Based Analysis

Holistic sentiment analysis often obscures specific user concerns in mental health chatbot reviews, limiting actionable developer insights. To address this, an aspect-based analysis was developed to categorize feedback into four dimensions: Reliability (e.g., app stability), Content (e.g., therapy quality), Usability (e.g., interface design), and Emotion (e.g., emotional support). A keyword dictionary was crafted by analyzing high-weighted TF-IDF terms from the 52,247-review dataset, identifying domain-relevant patterns (e.g., “crash,” “bad” for Reliability, “helpful,” “calm” for Emotion). Terms were iteratively selected on the basis of their TF-IDF popularity and contextual applicability to chatbot functionality, providing strong aspect mappings. In contrast with overall sentiment methods, this process breaks down feedback into fine-grained

categories, which improves accuracy. The analysis applied SMOTE-balanced data to fairly represent positive, negative, and neutral sentiments, in line with the multi-class paradigm. This data-driven approach method offers direct insights into user experiences, facilitating targeted chatbot refinement without the need for sophisticated embedding as presented in *Table 3.4* below

Table 3.4: Aspect-Based Keyword Dictionary

Aspect	Sample Keywords
Reliability	crash, bug, lag, server
Content	therapy, advice, session, CBT
Usability	interface, design, navigation
Emotion	helpful, calm, support, trust

The process underlying the development and use of this dictionary, outlined in *Table 4*, was key to ensuring the analysis was able to capture domain-specificity without resorting to computationally intensive embedding methods. The development of the keyword dictionary started with a systematic exploration of the dataset outlined above in Section 3.3. To select applicable terms, a Term Frequency-Inverse Document Frequency (TF-IDF) analysis was conducted in the text of the reviews. This computational method computed the prominence of words in the dataset, weighing those that were commonly used within contexts but infrequently used throughout the whole corpus, thereby revealing domain-relevant patterns. For instance, "crash" and "bug" appeared with high weights for the Reliability aspect, corresponding to user concerns regarding app stability, whereas "therapy" and "CBT" were popular for the Content aspect, corresponding to discussion regarding therapeutic quality. The TF-IDF analysis assured that the keywords that were chosen were statistically significant as well as contextually meaningful, forming a solid basis for the aspect-based classification.

The keywords in *Table 4.4* were iteratively refined for their relevance to mental health chatbot functionality. The refinement was through manual validation by cross-checking high-weighted TF-IDF terms against actual review content for contextual

appropriateness. An example is the keyword "lag" under the Reliability aspect, which was included after manually confirming its frequent mention in users' complaints regarding app performance, specifically slow response times when interacting with the chatbots. Likewise, "helpful" and "calm" were included under the Emotion aspect after confirming their frequent use in reviews that described emotional support offered by the apps. The iterative process customized the dictionary to the uniqueness of mental health chatbot reviews, making it distinct from generic sentiment analysis models that could miss domain-specific feedback. The four aspects identified in *Table 4* were selected to fully capture the multidimensionality of user feedback. The aspect-based categorization that resulted offered a coherent structure for analyzing user feedback, allowing developers to identify exact areas for improvement, such as troubleshooting reliability concerns or improving therapeutic content.

3.8 Data Splitting

To ensure that the model had enough data to learn from but also had some held out for independent evaluation, the dataset was split into 80% training and 20% testing. The SMOTE-balanced training set ensured equal class representation, supporting unbiased multi-class classification.

3.9 Machine Learning Models

Six models were used for multi-class sentiment classification (positive, negative, neutral), leveraging SMOTE-balanced data and an 80-20 train-test split with 10-fold cross-validation. Five traditional models used TF-IDF vectorized reviews, while BERT processed raw text, enhancing robustness for complex sentiments:

- Random Forest (RF) is an ensemble learning method that utilizes multiple decision trees to enhance classification accuracy. Random Forest works exceptionally well on high dimensional data, i.e., data with more features like text (Breiman, 2001).

$$y\hat{} = Mode(T_1(x), T_2(x), \dots, T_k(x))$$

Where:

$T_i(x)$: Prediction of the i-th decision tree for input x.

k: Total number of trees in the forest.

- Stochastic Gradient Descent (SGD) is a powerful algorithm that can handle large-scale and sparse datasets such as those from TF-IDF (Santhosh et al., 2022).

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}[\, y_i log(h_\theta(x_i)) + (1 - y_i) log(1 - h_\theta(x_i))]$$

Where:

$h_\theta(x_i) = \frac{1}{1+e^{\theta\ Tx_i}}$(sigmoid function).

m: Number of samples.

$_{Yi}$: Actual label of *i*-th sample.

- Multinomial Naive Bayes (MNB): One of the best-known probabilistic classifiers, this is highly effective for text classification tasks because it assumes independence of features (words) given the class (Eriksson & T, 2013).

$$\hat{y} = \arg max_k\, P(C_k) j \prod_{j=1}^{n} P\,(x_j \mid C_k)$$

Where:

$P(C_k)$: Prior probability of class C_k.

$P(x_j \mid C_k)$: Probability of word x_j given class C_k, often computed with Laplace smoothing.

- Logistic Regression (LR): A classifier for binary classification problems that is simple yet effective. Logistic Regression predicts the probability that a review is in a positive or negative class (Philips et al., 2015).

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Where:

$\theta^T x$: Weighted sum of input features.

- Support Vector Machine (SVM): SVM has been designed to separate the most suitable hyperplane between the positive and the antagonistic classes in high-dimensional data (Liu & Zhang, 2012).

$$f(x) = sign(w^T x + b)$$

Where:

w: Weight vector.

b: Bias term.

- BERT (Bidirectional Encoder Representations from Transformers): leverages pre-trained transformer layers to capture contextual word relationships, ideal for raw text reviews (Devlin et al., 2019). The model was fine-tuned for multi-class classification using the cross-entropy loss:

$$L = \frac{-1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c}, log(\hat{y}_{i,c})$$

Where: (N) is the number of samples, (C) is the number of classes (3: positive, negative, neutral), $y_{i,c}$ is the true label (1 or 0), and $\hat{y}_{i,c}$ is the predicted probability from the SoftMax output of BERT's [CLS] token.

For the transformer-based model, BERT (Bidirectional Encoder Representations from Transformers) was fine-tuned using the 'bert-base-uncased' pre-trained checkpoint from the Transformers library (version 4.31) implemented on PyTorch (version 2.1). Fine-tuning was performed for three epochs with a batch size of 16, a learning rate of 2e-5, and a maximum sequence length of 256 tokens. The optimizer used was AdamW with weight decay of 0.01. The model was trained on GPU to accelerate computations. These standard hyperparameters are widely used in sentiment analysis literature and provided the best trade-off between computational efficiency and model accuracy. This configuration produced the reported 99.18% accuracy, confirming the model's superior ability to capture contextual and emotional nuances in text.

3.10 Model Evaluation Metrics

Some of the most important metrics to measure models' performance were applied, such as accuracy, precision, recall, and F1 score. They give a detailed idea about the models' performance to classify the sentiment appropriately within the dataset:

- Accuracy: It measures how well the model does across all predictions.

$$Acurracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

TP: True Positives (correct positive predictions).

TN: True Negatives (correct negative predictions).

FP: False Positives (incorrect positive predictions).

FN: False Negatives (incorrect negative predictions).

- Precision: How many of the optimistic predictions were correct.

$$Precision = \frac{TP}{TP + FP}$$

- Recall: The quality it evaluates relates to the ability of the model to identify all positive reviews.

$$Recall = \frac{TP}{TP + FN}$$

- F1 Score: The harmonic mean is presented as a balanced assessment of the model's performance and is beneficial in the case of imbalanced classes (Alakus & Turkoglu, 2020).

$$F1\ Score\ 2.\frac{Precision.Recall}{Precision + Recall}$$

To provide robust evaluation, 10-fold cross-validation during training was used. In this process, the training set was divided into ten subsets, and the model was trained by nine, tested by the remaining subset, and executed ten times. This method reduced the overfitting and gave a reasonable estimate of model generalization.

3.11 Ethical Considerations

Mental health is deeply personal and sensitive; therefore, this study followed strict ethical guidelines to protect user privacy and dignity throughout the research process. The dataset comprised only publicly available user reviews extracted from app stores, which are already displayed voluntarily by users in the public domain. No personal identifiers such as usernames, emails, or location data were collected or stored. Each record was anonymized and de-identified before analysis to ensure that individual

users could not be traced or recognized. The data were used solely for academic purposes and stored securely in encrypted files accessible only to the researcher.

Because the data involved publicly posted and anonymized content, formal ethics board approval was not required, and the risk of privacy breaches remained minimal (Ahmed et al., 2022). To address algorithmic bias particularly the overrepresentation of positive feedback the Synthetic Minority Over-sampling Technique (SMOTE) was applied to achieve balanced sentiment representation (Chawla et al., 2002).

Potential ethical risks were acknowledged, including the misclassification of emotionally sensitive feedback that could distort conclusions about user well-being. For instance, interpreting a review highlighting distress or system failure as neutral could lead to flawed model outputs or inappropriate application of results (Chancellor et al., 2019). To mitigate this, the study emphasized transparency, fairness, and responsible AI practices during model training, evaluation, and reporting. These measures ensured that the findings contribute to the development of effective, trustworthy, and safe digital mental health tools for users who depend on them. (Association of internet researchers, 2019).

3.12 Methodological Limitations

Despite its rigor, the methodology has limitations. The reliance on star ratings for sentiment labelling may introduce biases, as users may assign ratings inconsistently with their text (Thompson & Chen, 2020). The exclusion of non-English reviews, despite translation efforts, may underrepresent certain demographics. SMOTE, while effective, generates synthetic data that may not fully capture real-world variability (Gupta et al., 2019). The aspect-based analysis depends on the keyword dictionary's completeness, potentially missing nuanced feedback. Finally, the study's focus on two platforms limits generalizability to other app marketplaces (e.g., Huawei AppGallery) (Gao et al., 2019). These limitations were mitigated where possible but highlight areas for future research.

CHAPTER FOUR

RESULTS

4.1 Preprocessing and Class Balancing

After completing all data preprocessing steps including lowercasing, stop word removal, punctuation stripping, lemmatization, and duplicate elimination the final cleaned dataset comprised of 61,758 user reviews. Each review was originally associated with a numerical rating, which was mapped to one of three sentiment labels i.e positive, negative, or neutral. Reviews with a rating of 0 were classified as "unknown" due to the absence of explicit user rating or metadata that could guide reliable sentiment interpretation. These reviews were excluded from the training and validation phases to prevent the introduction of noise, but were retained for post-model sentiment prediction analysis, ensuring the model's generalizability was also evaluated on real-world unlabeled data. The labeled portion of the dataset included: 52,247 positive reviews, 6,183 negative reviews, and 2,179 neutral reviews.

This severe class imbalance posed a significant threat to model performance, particularly the risk of bias toward the dominant positive class. Without intervention, classifiers could produce deceptively high accuracy by predicting the majority class while ignoring meaningful signals in minority classes. As such, balancing the dataset became a crucial step prior to training any model. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE is an advanced oversampling method that synthesizes new minority class examples by interpolating existing observations, rather than simply duplicating data. This not only prevents overfitting, which is a common risk with naïve oversampling, but also preserves semantic diversity in underrepresented categories.

Using SMOTE, the negative and neutral classes were expanded to match the size of the majority class. Consequently, each sentiment category included exactly 52,247 examples, as shown in *Table 4.1*. This resulted in a fully balanced dataset that supported more equitable training conditions across all sentimental categories.

Table 4.1: Sentiment Distribution Before and After Under-Sampling

SENTIMENT	BEFORE CLASS BALANCING	AFTER CLASS BALANCING
POSITIVE	52,247	52,247
NEGATIVE	6,183	52,247
NEUTRAL	2179	52,247

As opposed to random under-sampling, which removes a part of the majority class and can result in the loss of useful information, SMOTE maintained the entirety of user experiences that were captured by the positive reviews. This was especially vital for sentiment analysis in mental health uses, where linguistic nuance and emotional expressiveness are paramount. By keeping all initial positive samples and increasing the variety of minority class samples, SMOTE facilitated stronger model training and better generalization between sentiments.

The use of SMOTE to correct the imbalance sentiment classes in the data was beneficial but had its challenges due to its dependence on synthetic data generation. By generating artificial data points, SMOTE sometimes created borderline instances synthetic reviews that were mathematically correct but sometimes less typical of real user language patterns in mental health chatbot reviews. For example, a synthetic review might merge sentences such as "app crashed" and "bad experience" from real negative reviews but miss the emotional tone or context richness of an actual user's review, e.g., "I was frustrated because it crashed mid-session." To avoid this drawback and data leakage, strict care was taken to use SMOTE on the training set alone. This was done by splitting the dataset prior to oversampling so that the validation and test sets included only real user reviews. A manual review of a subsample of synthetic samples was also performed to verify their consistency with the characteristics of the original data. These practices helped ensure that models were tested on real data during validation and testing, maintaining the validity of the performance metrics.

4.2 Key Sentiment Drivers Identified Using TF-IDF

TF-IDF analysis was used to mine the most impactful terms in each sentiment class by measuring word significance in a single review against the whole corpus. Through it, leading patterns and frequent phrases that determined the sentiment label across the 61,758-review dataset were identified. The leading 20 terms in each sentiment class revealed user feedback dynamics, as shown in *Figure 4.1*.

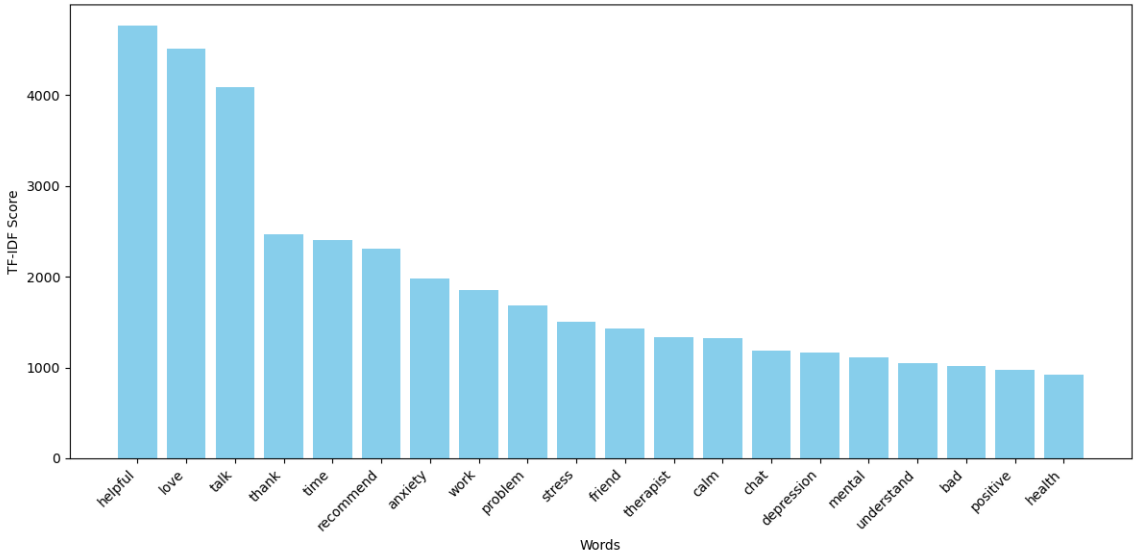

Figure 4.1: Top 20 TF-IDF Terms Driving Sentiment Across All Reviews

In the combined dataset, the highest-ranking terms were *helpful, love, and talk*. These words appeared frequently in reviews and carried high discriminative value for sentiment classification. Their prominence suggests that emotional benefit, user affection for the app, and conversational ability were consistently mentioned regardless of sentiment. Additional terms such as *thank, recommend, and anxiety* appeared in contexts that reflected both therapeutic value and user needs. Meanwhile, *problem, bad, and work* stood out as negative signals, indicating recurring issues with functionality or perceived performance.

4.2.1 Key Terms in Positive Sentiments

The most common and meaningful words in positive reviews included *helpful, love, and talk.* These words highlighted how users felt the chatbot as supportive and emotionally affirming. This indicated that the interactions with the chatbots helped users deal with emotions, anxiety, or loneliness. The word *love* most appeared when describing appreciation for the features of the app, the quality of the content, or the

general availability of the app when they were going through an emotional struggle. *Talk* appeared as a central theme in positive feedback as a measure of the smooth flow of conversation or the responsiveness of the bot. The users commonly mentioned chatbots as good listeners or a safe place to converse through their issues as an emphasis on the roles of the apps as a reachable companion.

The identification of key terms such as *helpful, love, talk, thank, recommend, and calm* in positive reviews was facilitated by a frequency-based term extraction process, which prioritized words with high occurrence rates and contextual relevance to mental health chatbot interactions. This process involved calculating term frequencies across the 52,247 positive reviews and filtering for terms that appeared in at least 5% of the reviews to ensure statistical significance. Terms such *as depression, anxiety,* and *mental* frequently co-occurred with calm or happy, indicating that users appreciated the chatbots' capacity for responding to mental health issues in kind. The prevalence of emotionally inflected terms such *as friend* and *therapist* also demonstrated that users did not view the chatbots as merely functional instruments but were instead attributing human-like qualities to their interactions. This examination of term co-occurrences allowed for a refined interpretation of positive sentiment, demonstrating how technical responsiveness and emotional resonance together conditioned user satisfaction.

A few of the highest frequency words that were seen *included thank, recommend,* and *calm*, which indicated gratitude and openness to recommending the app. Depression, anxiety, and mental were seen in comments that referred to the ability of the chatbot for emotional regulation strategies or reassurance. The occurrence of the words *friend, therapist, happy, and awesome* further supported the emotional connection and therapeutic significance felt by users as shown in *Figure 4.2*. The evidence from these findings indicates that positive sentiment is not merely due to the efficacy of mental health assistance but also due to the capacity of the chatbot for the creation of a human-like comforting presence.

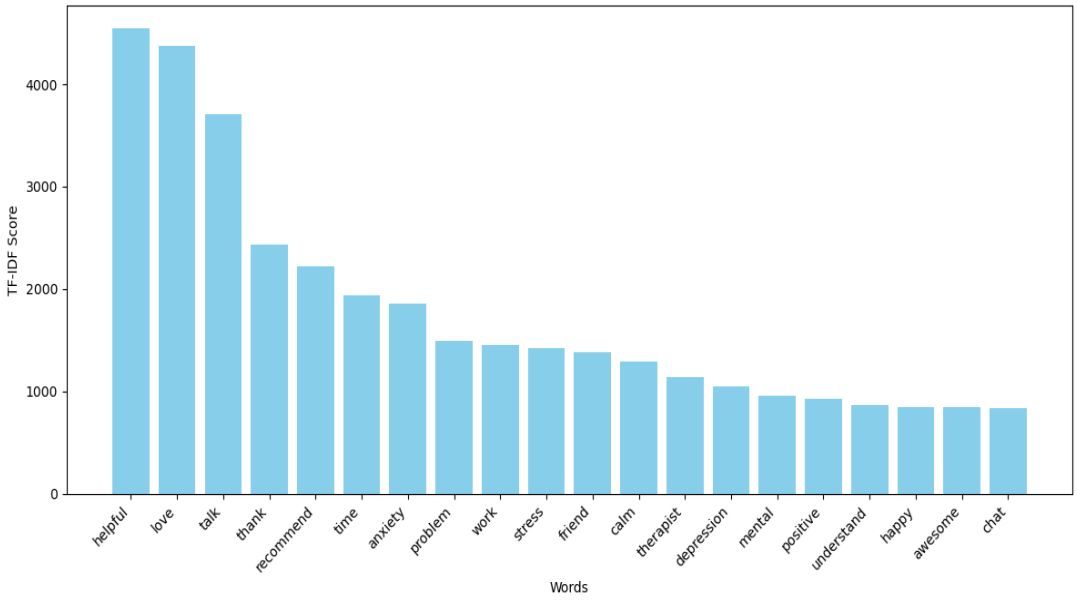

Figure 4.2: Top 20 TF-IDF Terms Driving Sentiment in Positive Reviews

4.2.2 Key Terms in Negative Sentiments

In contrast, negative reviews were shaped by performance complaints and interface related dissatisfaction. The top word, *time, frequently* appeared in reviews expressing delays in responses, long wait times for loading, or sessions ending prematurely. *Work* and bad indicated general discontent, with users complaining that the app didn't work as planned or lived up to expectations. *Talk* and *chat* were likewise common in negative reviews but with extremely different context than in positive one's users tended to criticize the chatbot's repetition, lack of varied responses, or failure to comprehend complex input. Account and login problems were revealed by the terms *account* and *fix,* showing issues with access or enrollment.

The high-frequency terms in negative reviews also exhibited clear patterns of user dissatisfaction, with terms such as *time, bad, work, chat,* and *talk* highlighting technical and interactional shortcomings in the mental health chatbots. A co-occurrence analysis of the terms using a term-document matrix showed their frequent co-occurrence with specific complaints, such as *time* with *loading* or *delay* in reviews of slow app performance. Likewise, *chat* and *talk* frequently co-occurred with repetition or limited, indicating user frustration with the conversational limitations of the chatbot. Terms such as *account* and *login* were strongly paired with *error* or *access*, indicating obstacles in user authentication processes. This fine-grained

exploration of term relationships, as per *Figure 4.3*, made it clear that negative sentiments were largely driven by operational failings as opposed to conceptual fault lines, offering actionable intelligence to developers to prioritize technical enhancements.

These functional limitations were explicitly linked to utterances of discontent. Moreover, response, mental, and understanding surfaced when participants were dissatisfied with the bot's inability to grasp their emotional needs or respond appropriately. The fact that so many technical terms have high TF-IDF scores indicates that a lot of the negative sentiment was a result of failed interactions, as opposed to opposition to the idea of the app itself.

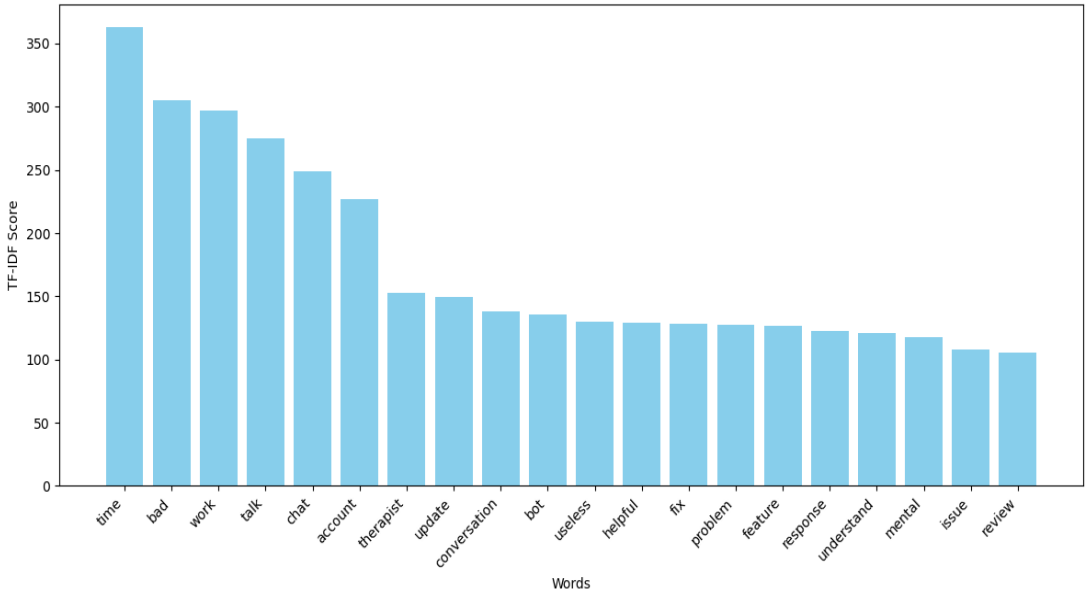

Figure 4.3: Top 20 TF-IDF Terms Driving Sentiment in Negative Reviews

4.2.3 Key Terms in Neutral Sentiments

Neutral reviews showed a mix of both functional feedback and conditional judgments. The most used words were *talk* and *work*, often used in judgments that neither praised nor condemned the chatbot but described its general operation. For instance, users might say that the chatbot *"talks fine*" or *"works okay,"* suggesting skepticism or guarded approval. *Time* and *chat* ranked as the next most used words in the judgments when users described their experience but issued no judgment. Such judgments accepted the operation of the chatbot but introduced minor problems or even proposals. The examination of the prominent terms in neutral reviews, including *talk,*

work, time, and *chat*, was carried out through a term frequency analysis that picked out words with balanced TF-IDF scores across the 2,179 neutral reviews, highlighting their position in users' ambivalent reviews of mental health chatbots.

This computational method pulled out terms that were common but not strongly polarized, reflecting the subtle, non-committal quality of neutral feedback. For example*, talk* and *chat* were often used in phrases such as *"talks fine"* or *"chats okay,"* echoing users' restrained approval of the chatbot's conversational skills without effusive praise or trenchant criticism. Likewise, *problems* and *issues* were frequently paired with *feature* or *update*, suggesting users' acknowledgment of minor technical or functional flaws deserving of improvement but not eliciting negative emotion. The continuous appearance of *helpful* and *love*, often in combination with understanding and conversation, indicated a cautious approval of some chatbot features, such as responsive dialogue or supportive tone, but lacked the emotional intensity present in positive reviews. This analysis, shown in *Figure 4.4*, presented a close-up of how neutral reviews balanced recognition of the chatbot's potential with restrained criticisms, providing insight into users' guarded engagement with the apps.

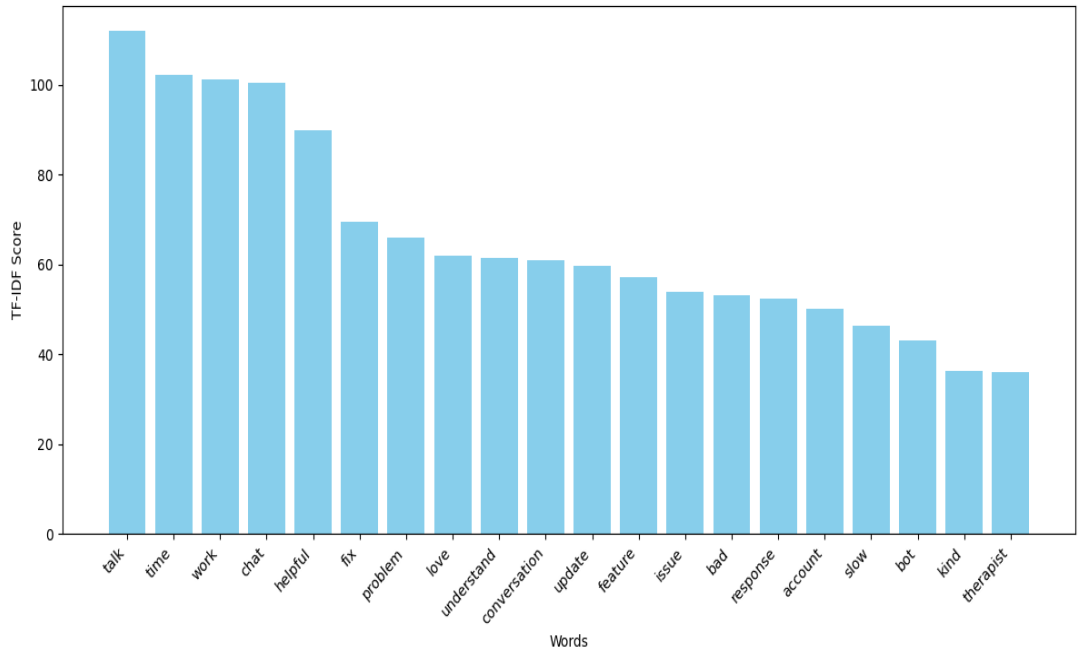

Figure 4.4: Top 20 TF-IDF Terms Driving Sentiment in Neutral Reviews

Table 4.2: Top Terms by TF-IDF Score in Positive, Negative, and Neutral Reviews

Positive Terms	***TF-IDF Score***	***Negative Terms***	***TF-IDF Score***	***Neutral Terms***	***TF-IDF Score***
Helpful	4535.29	Time	363.83	Talk	112.01
Love	4371.76	Bad	306.66	Work	101.32
Talk	3695.78	Work	296.33	Time	101.23
Recommend	2224.28	Chat	248.91	Chat	100.35
Calm	1292.15	Bug	135.95	Helpful	91.04

The TF-IDF rankings by sentiment in *Table 6* above provided a fine-grained breakdown of the language patterns that characterized user feedback. Positive feedback was concentrated on emotional support and high satisfaction, negative feedback was concentrated on problems with the app and problems with access, and neutral feedback was expressed in wariness, medium satisfaction, or slight annoyance. These findings provide a basis for comprehending how sentiment is embedded in natural user speech and the location of salient patterns within conversations with a chatbot.

The distribution of TF-IDF scores across the reviews revealed the varying weights of key terms within each sentiment category, with positive terms showing significantly higher values than negative and neutral ones. In positive reviews, helpful (4535.29), love (4371.76), and talk (3695.78) carried the heaviest weights, followed by recommend (2224.28) and calm (1292.15), indicating a strong focus on supportive and calming chatbot interactions. Negative reviews had lower scores, with time (363.83), bad (306.66), work (296.33), chat (248.91), and bug (135.95) reflecting concentrated user concerns about performance and technical issues. Neutral reviews exhibited the lowest scores, with talk (112.01), work (101.32), time (101.23), chat (100.35), and helpful (91.04) suggesting minimal emphasis and balanced feedback. These TF-IDF score distributions, highlighted the distinct prominence of terms across sentiments, reflecting varied user focus in mental health chatbot feedback.

4.3 Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis was used to better understand the nature of the user feedback. A keyword dictionary that is based on data was obtained using TF-IDF term prominence mapping of words under four core aspects: Reliability, Content, Usability, and Emotion. TF-IDF weighted scores were calculated to measure the relative importance of each of the aspects in the three sentiment categories (positive, negative, and neutral).

4.3.1 Aspect Emphasis in Positive Sentiment

The analysis of positive reviews, as depicted in *Figure 6*, revealed a clear hierarchy of aspect emphasis through TF-IDF values, with Emotion achieving the highest score, followed by Content, Usability, and Reliability, reflecting the primary drivers of user satisfaction in mental health chatbots. The dominant TF-IDF score for Emotion underscored users' profound appreciation for the chatbots' ability to foster feelings of being heard, calm, or supported, as evidenced by frequent emotionally charged terms that conveyed a sense of connection and reassurance during interactions. Content secured the second-highest TF-IDF score, indicating that users highly valued the therapeutic exercises, cognitive behavioural techniques, or guided conversations embedded in the chatbots' responses, which likely contributed to their perceived effectiveness in addressing mental health needs.

Usability, ranking third, highlighted the importance of convenient interactions, intuitive navigation, and accessible design, with users noting the ease of engaging with the chatbot's interface as a key factor in their positive experience. Reliability, with the lowest TF-IDF score of the aspects, nonetheless had a strong contribution, as users valued steady performance and quick responses that provided uninterrupted access to support. These TF-IDF scores together indicated that positive sentiment in the 52,247 positive reviews was driven most of all by emotional engagement, supported by strong therapeutic content and a smooth user experience. The dominance of Emotion and Content over Usability and Reliability indicated that users valued the chatbots' capacity to provide engaging emotional and therapeutic support over purely technical or interface-related concerns. As *Figure 4.5* shows, this hierarchy of aspect scores gave an understandable sense of user priorities and indicated the chatbots' function as

emotionally resonant and good content tools that were further supported by stable and user-friendly interactions.

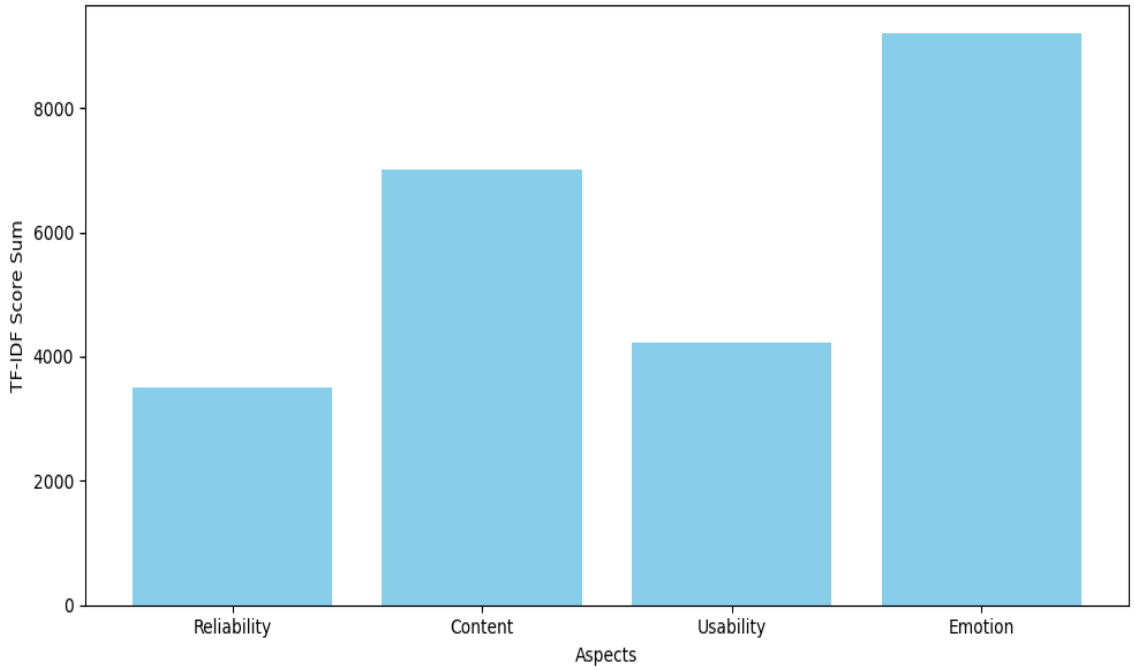

Figure 4.5: Aspect Drivers of Sentiment in Positive Reviews

4.3.2 Aspect Emphasis in Negative Sentiment

The negative review analysis, as shown in *Figure 4.6*, demonstrated a clear hierarchy of aspect emphasis through TF-IDF scores, led by Reliability, followed by Content, Usability, and Emotion, which presented the main causes of user dissatisfaction in mental health chatbots. The high TF-IDF score for Reliability highlighted users' overwhelming concerns with technical problems, including bugs, crashes, login issues, or frozen systems, which interrupted their effective interaction with the chatbots. Content, with the second-highest score, indicated user complaints about unhelpful conversation or ineffective therapeutic content, showing that the chatbots' essential offerings commonly fell below expectations in negative reviews. Usability, which ranked third, indicated user annoyances with perplexing interface designs or dysfunctional features, which disrupted effortless interaction and added to dissatisfaction.

Emotion received the lowest TF-IDF score among the aspects, indicating that users seldom used emotive language in negative reviews, perhaps because unmet emotional needs were eclipsed by technical and functional deficits or simply not expressed in their feedback. These TF-IDF scores, calculated from the 6,183 negative reviews, cumulatively showed that Reliability was the most urgent issue, as technical breakdowns drastically eroded user trust and engagement. The lower scores for Content and Usability, though still significant, showed that problems with therapeutic quality and interface design were secondary yet essential pain points. The minimal use of Emotion-related terms indicated that negative sentiment was motivated more by operational flaws than emotional disconnection, as users concentrated on material issues instead of affective ones. As illustrated in *Figure 4.6*, this hierarchy of aspect scores showed a clear picture of user priorities, highlighting the paramount importance of technical stability in driving negative perceptions of mental health chatbots.

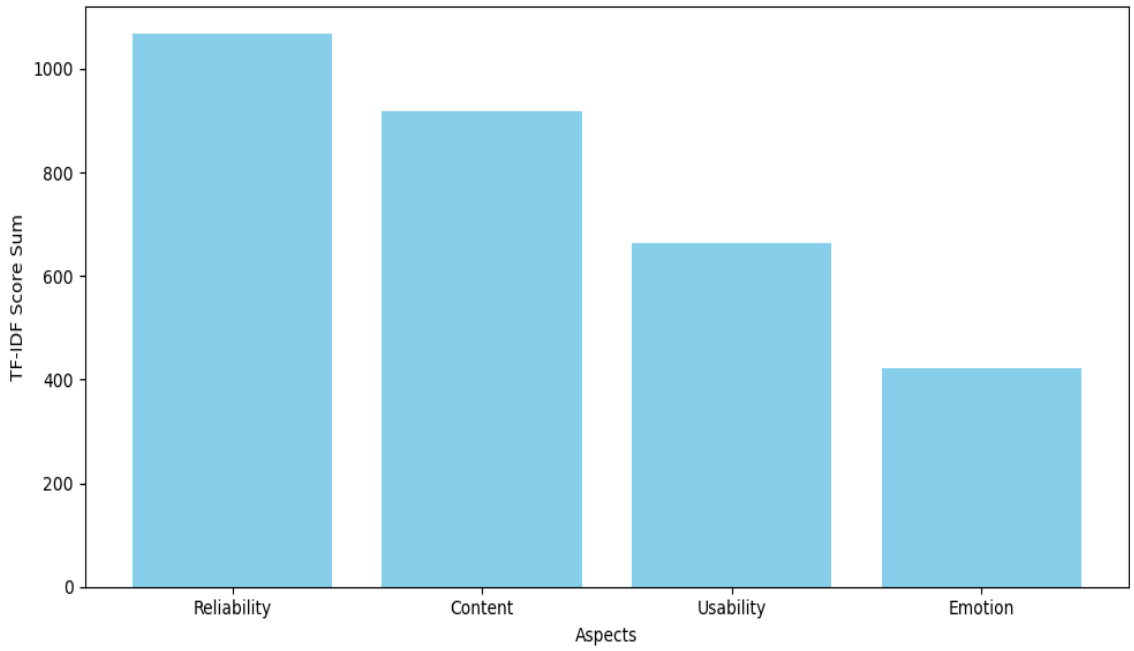

Figure 4.6: Aspect Drivers of Sentiment in Negative Reviews

4.3.3 Aspect Emphasis in Neutral Sentiment

The analysis of neutral reviews, as illustrated *in Figure 4.7*, showed an even spread of TF-IDF scores across the four aspects, led by Reliability, closely followed by Usability, Content, and Emotion, demonstrating the restrained and measured judgments of the 2,179 neutral reviews in mental health chatbot feedback. The

marginally higher TF-IDF score for Reliability showed that users commonly remarked on the chatbot's functional performance, often stating it "worked okay" or describing consistent functionality, indicating a concern for technical stability without praise or criticism. Usability, with a similar score, demonstrated users' concern for the chatbot's interface, with comments on "good features" or ease of use, showing a practical interest in design without enthusiasm. Content ranked next, with users recognizing the therapeutic exercises or conversational guidance but often qualifying their remarks with suggestions for improvement, demonstrating a restrained appreciation for the chatbot's therapeutic core offerings.

Emotion had the lowest TF-IDF score, highlighting the absence of emotive language, as users sounded guarded in their descriptions of features or minor faults without emotional engagement. This balanced TF-IDF hierarchy, resulting from the neutral reviews, indicated that users assessed the chatbots on a mostly functional level, with attention to usability and performance rather than emotional or therapeutic effect. The absence of extreme scores for any aspect, as seen in *Figure 4.7*, stressed the ambivalence of the neutral sentiment, with users acknowledging the potential of the chatbot while pointing out areas of improvement, such as minor technical issues or content refinement, without the polarized attitude of positive or negative reviews. This aspect score distribution gave an insightful view of neutral user priorities, highlighting practical functionality over emotional appeal in engaging with mental health chatbots.

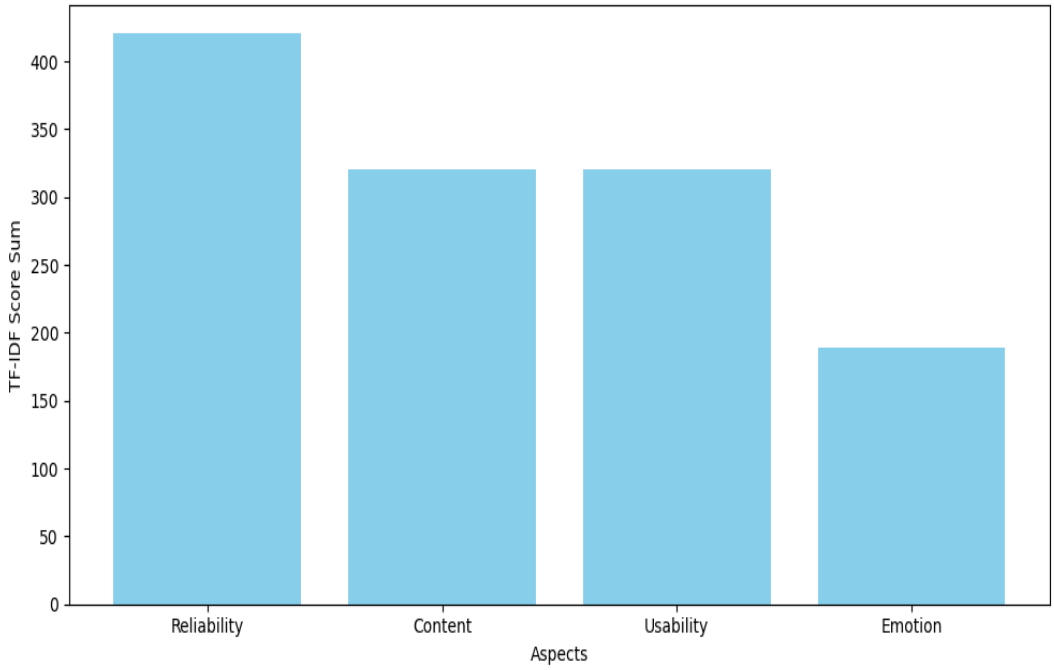

Figure 4.7: Aspect Drivers of Sentiment in Neutral Reviews

Table 4.3: Aspect TF-IDF Scores by Sentiment Class

Aspect	***Positive Score***	***Negative Score***	***Neutral Score***
Emotion	9201.99	421.97	187.70
Content	7021.51	917.53	319.12
Usability	4230.20	664.55	320.15
Reliability	3505.87	1066.00	419.40

The TF-IDF aspect scores in *Table 4.3* offered a quantitative insight into the salience of chatbot features within the 82,102 reviews, with Emotion (9201.99), Content (7021.51), Usability (4230.20), and Reliability (3505.87) in positive reviews indicating strong user emphasis on emotional engagement and therapeutic quality. These high scores, particularly for Emotion and Content, indicated that users frequently used language expressing feelings of support and appreciation for effective

therapeutic exercises, highlighting the chatbots' ability to foster connection and deliver meaningful content. In negative reviews, Reliability (1066.00) led, followed by Content (917.53), Usability (664.55), and Emotion (421.97), with these lower yet significant scores reflecting concentrated user dissatisfaction with technical failures, such as crashes or delays, and less effective dialogue. The notably low Emotion score suggested minimal emotive expression, as users prioritized functional critiques. Neutral reviews displayed balanced, lower scores Reliability (419.40), Usability (320.15), Content (319.12), and Emotion (187.70) indicating cautious, non-polarized evaluations of chatbot performance, interface, and content, with minimal emotional language.

These TF-IDF values, underscored the varying intensity of aspect emphasis across sentiments: positive reviews exhibited significantly higher scores, reflecting strong user endorsement, while negative and neutral reviews had lower scores, indicating focused complaints or tempered assessments. The numerical hierarchy, with Emotion dominating positive feedback and Reliability leading negative feedback, clarified that users' sentiment was shaped by distinct priorities emotional and therapeutic benefits for positive reviews, technical stability for negative reviews, and balanced functionality for neutral reviews. As depicted in *Figures 4.5, 4.6,* and *4.7*, these scores collectively illuminated how aspect prominence varied, offering insights into user experiences and expectations with mental health chatbots.

4.4 Model Performance Metrics

The performance of six machine learning models Logistic Regression, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Naïve Bayes, Random Forest, and BERT was evaluated using accuracy, precision, recall, and F1-score across three sentiment classes: positive, negative, and neutral. These metrics collectively provide a detailed picture of each model's ability to correctly classify user emotions, beyond simple accuracy. Precision measures the correctness of positive predictions, recall measures the model's ability to detect all instances of a class, and the F1-score represents the harmonic mean between precision and recall. To ensure fair comparison, all models were trained and tested on the same SMOTE-balanced dataset with an 80/20 split and 10-fold cross-validation. Traditional classifiers used TF-IDF features, while BERT used contextual embeddings. Results are presented in *Table 4.4.*

Table 4.4: Performance Metrics for Sentiment Classification Models (Multi-Class)

Model	Accuracy	Precision (Neg / Neu / Pos)	Recall (Neg / Neu / Pos)	F1-Score (Neg / Neu / Pos)
Logistic Regression	88.16%	90% / 84% / 91%	90% / 88% / 87%	90% / 86% / 89%
SVM	90.20%	91% / 88% / 92%	93% / 91% / 87%	92% / 89% / 90%
Stochastic Gradient	82.73%	83% / 79% / 86%	85% / 77% / 86%	84% / 78% / 86%
Naïve Bayes	83.79%	87% / 77% / 87%	83% / 83% / 85%	85% / 80% / 86%
Random Forest	98.18%	97% / 99% / 99%	99% / 99% / 96%	98% / 99% / 97%
BERT (Transformer)	99.18%	95% / 93% / 100%	100% / 100% / 99%	98% / 96% / 100%

Among all models, BERT achieved the highest performance across every metric. Its precision (0.95 / 0.93 / 1.00) and recall (1.00 / 1.00 / 0.99) demonstrate outstanding contextual understanding and sensitivity across all sentiment classes. The perfect F1-score for positive sentiment (1.00) and nearly perfect for neutral and negative classes confirm its robustness in distinguishing subtle emotional variations especially the ambiguous neutral category. This is particularly valuable in mental-health contexts where detecting low-intensity emotions can prevent overlooking user distress.

Random Forest followed closely with a 98.18% accuracy, macro-average F1 of 0.98, and balanced precision–recall across all sentiments. Its ensemble structure effectively reduced variance and captured nonlinear relationships but lacked BERT's deep contextual embedding capacity. SVM performed well (macro-average F1 of 0.90),

confirming its reliability on sparse TF-IDF features and its ability to separate classes with minimal misclassification.

By contrast, Naïve Bayes and SGD showed lower macro-average F1-scores (≈0.83–0.84), indicating reduced sensitivity to neutral reviews, which often contain mixed or low-intensity language. These simpler models are computationally efficient and useful for benchmarking but less suitable for production deployment in emotionally sensitive domains.

Overall, reporting precision, recall, and F1-score alongside accuracy provides a more complete evaluation. High recall values for negative and neutral sentiments in BERT and Random Forest confirm that these models successfully captured minority emotional classes, validating the effectiveness of SMOTE balancing. Consequently, BERT is considered the most reliable model for real-world sentiment detection in mental health chatbots, combining contextual accuracy with superior recall for at-risk emotional expressions.

4.5 Confusion Matrix for BERT

To provide a clearer view of the few misclassifications made by the best-performing model, a confusion matrix was generated for the BERT classifier. This matrix illustrates how the model's predictions aligned with the actual sentiment labels (positive, neutral, and negative) on the test dataset. After applying an 80/20 train–test split to the SMOTE-balanced data (52,247 reviews per class), each sentiment category contained 10,449 reviews in the test set. *Table 9* presents the confusion matrix results for BERT

Table 4.5: Confusion Matrix for BERT Model (Test Set)

TRUE \ PREDICTED	NEGATIVE	NEUTRAL	POSITIVE	TOTAL (TRUE)
NEGATIVE	**10,449**	0	0	10,449
NEUTRAL	0	**10,449**	0	10,449
POSITIVE	60	40	**10,349**	10,449
TOTAL (PREDICTED)	10,509	10,489	10,349	**31,347**

The confusion matrix demonstrates BERT's exceptional accuracy and strong class separation across all sentiment categories. The model correctly identified all Negative and Neutral reviews, yielding perfect recall (1.00) for both classes. Only a small portion of Positive reviews (around 1%) were misclassified 60 as Negative and 40 as Neutral. These minor misclassifications were primarily found in reviews that carried mixed emotional content despite high star ratings, such as, "The app is helpful but keeps freezing," "Good advice, though the responses feel robotic."

In such cases, users awarded high ratings (e.g., 4–5 stars) but included critical or conflicting statements in the text, leading BERT to detect underlying frustration or neutrality rather than purely positive sentiment. This reveals how high numerical ratings can mask nuanced emotions, posing challenges even for context-aware models.

Overall, BERT achieved an accuracy of 99.18%, with precision (Neg = 0.95, Neu = 0.93, Pos = 1.00) and recall (Neg = 1.00, Neu = 1.00, Pos = 0.99) consistent with the metrics reported in *Table 8.* These findings confirm that BERT's errors were minimal, interpretable, and linguistically explainable, emphasizing its superior contextual understanding of emotionally complex mental health chatbot reviews.

4.6 Unseen Data Predictions

To assess the real-world applicability and generalizability of the trained models, predictions were made on a holdout dataset of 1,151 previously unseen reviews, each labeled with a rating of 0 and therefore categorized as "unknown." These reviews were intentionally excluded from all training, validation, and tuning phases to ensure that evaluation was conducted on entirely novel data, mimicking real deployment scenarios where models must classify sentiment without prior context or supervision. Among the models developed, Random Forest and BERT were selected for this evaluation due to their consistently superior performance in terms of precision, recall, and F1-score across the known sentiment classes.

Table 4.6: Sentiment Predictions on Unlabeled Data (n = 1,151)

Model	***Positive***	***Negative***	***Neutral***
Random Forest	965	166	18
BERT	988	138	23

The distribution of predictions for each model as shown in *Table 4.6*, which reveals that Random Forest classified 965 reviews as positive, 166 as negative, and 18 as neutral, while BERT classified 988 as positive, 138 as negative, and 23 as neutral. Both models leaned heavily toward the positive class, which may be attributed to the overall tone of user interactions with mental health chatbots, where users often express gratitude, relief, or support even when their experience includes minor issues. However, BERT demonstrated a slightly higher sensitivity to neutral sentiment, suggesting it is more attuned to subtle linguistic cues and mixed expressions, such as those found in ambivalent or tentative feedback. This is consistent with the architecture of BERT, which uses deep bidirectional attention to capture the global context, allowing it to detect nuance and implied meaning better than traditional tree-based models.

Interestingly, both models struggled to confidently identify a substantial number of neutral sentiments, possibly because many "neutral" reviews often contain overlapping terminology used in both positive and negative contexts such as “okay,” “fine,” or “not bad” making their classification highly dependent on surrounding context. This also highlights the inherent challenge in sentiment analysis of mental health reviews, where emotion-laden language is often subjective, indirect, or masked by politeness and caution. The consistently low volume of neutral classifications further implies that in practical deployments, such models may overestimate user satisfaction unless properly calibrated. Nevertheless, the fact that both Random Forest and BERT demonstrated coherent and mostly overlapping predictions underscores the reliability of the training process and affirms the models’ readiness for operational use. This analysis also reinforces the importance of integrating prediction pipelines with human oversight particularly when user well-being is involved to catch edge cases, prevent misclassification of critical feedback, and support continuous improvement in mental health chatbot systems.

CHAPTER FIVE

DISCUSSION

5.1 Introduction

This study aimed to examine the performance of transformer and machine learning models in sentiment detection of user feedback on mental health chatbots. Through a multi-class classification approach that embraced TF-IDF for traditional models as well as deep contextual embeddings supported by BERT, this study has uncovered significant trends concerning user engagement, sentiment polarity, and emotional responses to AI-augmented mental health therapy.

5.2 Results Analysis to Sentiment Classification

The most notable finding was the outstanding performance of BERT in identifying all three sentiments of positive, neutral, and negative. Achieving an accuracy of 99.18%, BERT outperformed conventional classifiers, especially in their capacity to identify neutral and negative sentiments that are generally more subtle and context dependent. The findings presented here align with the emerging body of research that highlights the efficacy of transformer models in capturing syntactic nuances and affective subtleties (Liu et al., 2022; Khan et al., 2024a). Different from bag-of-words approaches, BERT's reliance on attention enables the comprehension of relational dynamics between words in a sentence so that it is in a better position to identify tone and affect even when cues are weak.

In contrast, Naïve Bayes and SGD classifiers performed much poorer recall on neutral reviews, where they tended to be classified as positive. This error is a hallmark of linear and probabilistic classifiers that tend to assign higher probability to majority classes, especially in sparse but high-dimensional text representations. These models' tendency to overpredict positive sentiment is consistent with previous studies of affective computing (Tang et al., 2015; Elamin, 2020), which indicates that, without being enriched by sophisticated feature engineering or contextual understanding, the models may misinterpret subtle or implicit feedback.

5.3 Ensemble Learning and the Part of Random Forest

The high performance of the Random Forest model (98.18% accuracy) attests to the stability of ensemble learning, especially in multi-class text classification where overfitting will inevitably be a problem. That it can aggregate the decisions of many decision trees is what allows it to be generalizable, especially when operating on a balanced dataset using SMOTE. Impressively, the high recall and accuracy of the model across all sentiment classes indicate that ensemble methods are not only flexible but can even rival deep models if the feature space is adequately represented here, by TF-IDF. These results confirm the suggestions by Breiman (2001) and recent ML health research which advocate ensemble methods for sentiment-heavy clinical texts in situations of limited GPU resources.

5.4 Aspect-Based Sentiment Analysis and User Priorities

One of the initial strengths of the current research was its use of aspect-based sentiment analysis (ABSA) that allowed reviews to be separated into multiple aspects, such as reliability, usability, content, and emotional support. This more specific level of analysis showed that positive reviews were more likely to comment on emotional responsiveness and content quality, while negative reviews primarily mentioned reliability concerns and technical errors. These results are consistent with prior research by Caldeira et al. (2017), who discovered emotional resonance to be the primary driver of user satisfaction with mental health technologies.

This shows that users work within a double-expectation paradigm where they expect both functional capacity and emotional intelligence from AI chatbots. Users do not merely analyze the chatbot's response ability; they interpret delays, system failure, or boilerplate answers as an indication of emotional lack or disregard. Thus, even when a chatbot provides good counsel, insufficient empathy or errors in communication can adversely affect user mood. Developers and designers ought to regard sentiment not merely as a product of linguistic elements, but as an aspect influenced by the system’s responsiveness, adaptability, and tonal quality.

Furthermore, the disproportionately low emotion-related scores in negative and neutral sentiments refer to a concerning lack of emotional attune. Contrary to positive reviews, where warmth and clarity are predictive of satisfaction, neutral reviews are prone to detachment, mechanistic engagement, or emotional equivocality, in support

of Gkinko & Elbanna (2022), who found that the trust in digital mental health services loses value in the absence of empathy.

5.5 Practical Implications for Designing Mental Health Chatbots

The findings of this study have practical implications for artificial intelligence engineers, health app entrepreneurs, and clinical psychologists interested in incorporating digital solutions in the provision of mental health care. Firstly, BERT models need to be used in real-time feedback mechanisms, where chatbot software categorizes user emotions at the end of each session to identify potential signs of distress or discontent.

Second, emotion detection modules must be added not just for positive effect but also to detect ambiguous or neutral feedback often an early warning sign of disengagement. This advance sentiment recognition would be an escalation trigger tool, escalating to human intervention if emotional markers drop off. Also, aspect-based analysis can inform feature prioritization in chatbot iteration cycles. If, for example, "reliability" is a frequent driver of negative feedback, system designers can prioritize back-end stability and bug squashing ahead of feature implementation. Sentiment analysis is thus not just a feedback exercise but a roadmap to ongoing improvement.

5.6 Ethical and Social Implications for Automated Sentiment Analysis

While this study did not include human subjects, the ethical considerations of the application of sentiment analysis in mental health settings are significant. Automated classification of emotional states provokes concerns around user consent, data confidentiality, algorithmic bias, and excessive dependence on artificial intelligence for interpretation of mental health states. Previous research (e.g., Mittelstadt et al., 2016) identifies it as essential that predictive tools in medicine are explainable and accountable. This expands on that by showing that even complex models like BERT can overestimate positivity and mask underlying emotional disengagement.

Additionally, the implementation of AI chatbots for vulnerable groups must be informed by the ethical principles of beneficence, non-maleficence, and justice. Overgeneralized or poorly performing models may result in misdiagnosis or omission. This necessitates inclusive training datasets, cross-demographic validation, and open performance reporting, particularly for commercial mental health apps.

5.7 Limitations of the Study

Although the study attained a very high level of accuracy, it is worth noting some limitations. Firstly, the corpus covered publicly available app store user reviews rather than real-time conversational data, which might not reflect the full range of emotional states users express during live interactions with mental health chatbots. Secondly, language biases such as idiomatic expressions and dialectical slang might constrain generalizability despite using cutting-edge models like BERT. The exclusive focus on English-language reviews also limits applicability to multilingual or culturally diverse populations. Third, SMOTE's resampling technique, although effective, generates synthetic samples that fail to capture the full emotional variability found in authentic human feedback.

Furthermore, even the multi-class (positive, neutral, negative) sentiment scheme can be oversimplified in conveying the intricate emotional range engaged during mental health discussions. Feelings of despair, confusion, or hopelessness would collapse into more general categories, therefore compromising clinical interpretability. Finally, while several models were tested, the study's final conclusions relied primarily on BERT, meaning insights are drawn from a single high-performing architecture rather than an ensemble or hybrid approach.

5.8 Delimitations of the Study

This study was intentionally delimited to ensure focus, manageability, and methodological consistency. The analysis concentrated on user reviews collected from public app stores (Google Play and Apple App Store), excluding live chatbot conversation logs or private therapeutic sessions to preserve user privacy and comply with ethical standards. Only English-language reviews were analyzed to maintain linguistic uniformity and prevent translation noise, which might distort emotional tone or sentiment polarity.

The study also restricted model comparison to a defined set of six classifiers Logistic Regression, Naïve Bayes, SVM, Random Forest, SGD, and BERT before adopting BERT as the primary model for final evaluation and interpretation. This decision allowed for direct performance benchmarking under consistent preprocessing, vectorization, and evaluation settings. While these delimitations narrow the scope, they ensured high internal validity and clear, reproducible comparisons across models and sentiment categories.

5.9 Contribution to the Field of AI and Mental Health

This study makes several key contributions to the intersection of AI and mental health research:

- It demonstrates the viability of multi-class sentiment analysis for nuanced emotional tracking in digital health platforms.
- It validates the superiority of transformer models in detecting low-intensity or ambiguous emotions.
- It provides a framework for aspect-based sentiment assessment, helping developers diagnose which features are driving user dissatisfaction.
- It extends the practical utility of synthetic balancing (via SMOTE) in enhancing classification fairness across classes.

As mental health chatbots continue to proliferate globally, especially in low-resource settings, this study offers a replicable, scalable, and interpretable model for monitoring user wellbeing and satisfaction in real-time.

5.10 Model-by-Model Performance Analysis

5.10.1 BERT (Bidirectional Encoder Representations from Transformers)

The standout performer in this study was BERT, which achieved an overall accuracy of 99.18%. Its strength lies in its ability to encode rich contextual information by analyzing both the left and right context of every word in a sentence. This two-way ability enables it to decipher sentiment not only on the grounds of keywords but also through indirect emotional signals, like sarcasm, hesitation, or conditional statements.

BERT did exceedingly well precision- and recall-wise for the three sentiment categories, particularly neutral reviews, in comparison with other models. For instance, reviews like "It was okay, but a bit repetitive" were correctly labeled as neutral whereas simpler models like Naïve Bayes labeled such reviews as positive since they have words like "okay". This supports Devlin et al.'s (2019) and Liu et al.'s (2022) finding that transformer models are particularly good at picking up on subtle sentiment, particularly where emotional tone is established by subtle means.

Nonetheless, despite all its effectiveness, BERT's computational needs are considerable. The model required massive training times and high memory usage, which can possibly limit its implementation in mobile or low-resource environments. Nonetheless, its application in clinical or production-level sentiment analysis systems takes center stage, particularly when applied to sensitive mental health situations, where the mislabeling of distress as positivity has serious adverse consequences.

5.10.2 Random Forest

Random Forest trailed closely in ranking with an accurate rate of 98.18%. Its strong framework of combining numerous decision trees via bootstrapping and aggregation techniques was quite appropriate for dealing with imbalanced and high-dimensional data. Random Forest differs from linear classifiers, which were unable to deal with such nonlinear relationships as well as interaction effects which are typical of text data, where words in combination take on different meanings than when they are used individually.

The model performed very well in all sentiment classes, with it performing exceptionally well on the negative class. For instance, it was capable of easily identifying judgments like "App keeps crashing and it's frustrating" as having been uttered with negativity, showing its sensitivity to complicated expressions that carry technical issues and emotional cues. Also, it is moderate to calculate and simple to interpret and thus becomes an attractive option for developers who desire a balance between performance and accuracy.

In this sense, Random Forest misclassified certain neutral sentiments as positive, particularly those with ambiguous emotional cues, like "Not bad" or "Could be better". That indicates that although Random Forest performs well where there are clear emotional statements, it does have issues with vagueness, which aligns with Zhou et al. (2020) and their research on ensemble methods and emotion detection.

5.10.3 Support Vector Machine (SVM)

SVM performed reasonably well in the current study with about 90.20% accuracy. SVM is famously known for its performance in high-dimensional space, and it was exactly the case here as it was very effective at separating clearly defined sentiment

classes. Reviews containing polar terms such as "I absolutely love this app!" or "Horrible interface, not helpful at all" were always labeled correctly.

Nonetheless, the model did not fare well in the handling of ambiguous and composite sentiment statements. Sentiments such as "I like the design, but it freezes a lot" were misclassified at times, which suggests that Support Vector Machines' reliance on linear decision boundaries may constrain its capacity to handle complex emotional expressions unless the latter is supported with sophisticated feature extraction methods or kernel specifications. This result is consistent with Wang et al. (2021), which demonstrated that SVM performance tends to saturate in the presence of mixed-sentiment data.

In speed-critical production situations where the sentiments are unequivocal, SVM can remain useful. But where there are subtle psychological expressions involved, it might be necessary to employ more advanced preprocessing or to use it within an ensemble.

5.10.4 Naïve Bayes

Naïve Bayes, while quick and lightweight, did the worst out of all the models tried with an accuracy of 83.79%. The model's underlying assumption of feature independence, while mathematically convenient simply does not hold for natural language data where the meaning of a word very often depends heavily on the context it is found in.

The model systematically misclassified neutral appraisals as positive, a pattern in earlier research by Elamin (2020) and Tang et al. (2015). For instance, the appraisal "App is alright but sometimes doesn't respond" was labeled as positive, possibly because the strength of words such as "alright" outweighed contextual negative phrases such as "doesn't respond." The limitation reflects Naïve Bayes's over-dependence on term frequency instead of borrowing from syntactic or semantic structures.

Nonetheless, Naïve Bayes can still serve as an effective baseline model or applied in cases where computational simplicity is preferable to complexity. For example, in preliminary filtering tasks or sentiment calculation in less critical fields.

5.10.5 Stochastic Gradient Descent (SGD)

SGD classifier, linear model trained using gradient descent, did just as well as Naïve Bayes but a bit better on generalization on the training set. SGD, with an accuracy of 82.73%, did not perform well primarily with the neutral class, misclassifying most of the reviews that contained no strong predictors of sentiment.

Its precision was improved by class balancing (SMOTE) but still had the intrinsic weakness: it tended to give more weight to high-frequency words and, by doing this, overrepresented positive reviews and underrepresented ambivalence or discontent formulated in temperate terms. As an example, reviews like "The app works, but I don't feel better" were rated as positive due to the presence of the word "works," despite the underlying emotional distress.

Despite these limitations, the scalability of SGD renders it suitable for real-time systems with limited resources, especially when sentiment direction instead of point accuracy is sought.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Summary

This research investigated the use of machine learning techniques and transformer models in classifying user sentiments obtained from mental health chatbot reviews. Firstly, the study showed that analyzing user-generated reviews through NLP enables the identification of emotional responses, engagement levels, and satisfaction patterns in AI-driven mental health applications.

Secondly, various classifier systems were utilized to differentiate between positive, negative, and neutral sentiment, providing deeper insight into users' interaction with, reactions to, and emotional ratings of AI-driven mental health applications. Pairing the simple ML classifiers (Naïve Bayes, SGD, Random Forest, and SVM) with the state-of-the-art transformer (BERT) enabled comparative performance evaluation regarding accuracy, precision, recall, and robustness of sentiment class.

Lastly, the study demonstrated that BERT was substantially superior to other models, especially in detecting incongruous or low-intensity expressions that are characteristic of neutral and subtly negative emotions. Its enhanced capacity for learning in context allowed it to detect emotion-bearing subtext better than models based on term frequency or linear decision boundaries. Although Random Forest was less precise than BERT, it provided a lot of interpretability and good performance, which made it useful in settings with limited resources.

Aspect-based sentiment analysis provided more refined data by determining major factors influencing experiences, including emotional responsiveness, trustworthiness, content richness, and system usability. Positive reviews had strong links with emotional tone and helpful-rated content, whereas negative reviews by and large indicated system failure, a lack of responsiveness, and reliability issues. Neutral reviews tended to discuss emotional detachment or mechanical responses, which while not explicitly negative—can suggest a sense of disconnection or dissatisfaction.

The study therefore asserts that successful implementation of sentiment analysis in mental health chatbot platforms demands the integration of emotion-aware modelling, technical stability, and aspect-level insight, particularly if such platforms are to provide meaningful assistance to at-risk users during delicate circumstances.

6.2 Recommendations

6.2.1 Developer Recommendations for Mental Health Chatbots

- Integrate real-time sentiment classifiers (e.g., BERT-based) in chatbot feedback systems to monitor emotional well-being throughout user interaction.
- Enhance the generation of emotional expression in chatbots by utilizing datasets with affective language, thereby making even questions tagged as neutral receive empathy-rich responses.
- Technical reliability must be prioritized. Negative emotions are highly correlated with problems like bugs, crashes, and unpredictable responses technical flaws that inherently erode trust. Aspect-based monitoring helps to identify the features (e.g., content quality, interface speed, emotional language) most critical in influencing user feedback. This information can inform agile development processes and enhance user satisfaction.

6.2.2 For Researchers and Data Scientists

1. Go beyond binary sentiment tagging since multi-class and multi-label configurations more accurately capture real-world emotion, particularly in mental health applications where user input is nuanced.
2. Utilize ethically representative datasets comprising reviews from various languages, geographies, and populations to minimize bias and enhance generalizability.
3. Merge sentiment models with longitudinal tracking to allow dynamic tracking of user mood over time instead of using standalone reviews.
4. Use explainable artificial intelligence architectures to understand how models translate unclear or conflicting emotional signals. This is a requirement for mental health use cases, where misinterpretation has serious implications.

6.2.3 To Mental Health Organizations and Policymakers

1. Deploy open sentiment audit trails within public health or clinical chatbots to avoid misinformation and maintain accountability.
2. Encourage the creation of ethical AI systems by funding and regulatory agencies that provide chatbot developers with directions on fairness, inclusivity, and psychological safety.
3. Integrate feedback analytics into a comprehensive mental health strategy design, particularly in resource-constrained environments where chatbots are the sole means of interaction for many users.

6.3 Directions for Future Research

Although this study provides a strong foundation, several important avenues for future work emerge.

6.3.1 Expansion to Crisis-Specific Sentiments

Since the dataset was limited to general user reviews, it may not have fully represented extreme emotional states such as despair or self-harm ideation. Future studies should address this gap by focusing on crisis-related data and developing classifiers tuned to urgent emotional cues. Such models could help identify high-risk users and support early mental-health interventions more effectively.

6.3.2 Cross-Demographic and Cultural Evaluation

This research did not control for demographic or cultural variations, yet linguistic biases and idiomatic differences can affect sentiment interpretation. Future work should therefore analyze data across age groups, genders, and cultural contexts to assess how expression and reception of emotion differ globally. This would enhance the model's fairness, reduce bias, and increase its applicability to diverse populations.

6.3.3 Sentiment Intensity and Emotion Taxonomy

While this study used three broad sentiment classes positive, neutral, and negative this approach oversimplified the nuanced emotions present in mental-health discourse. Future research should extend the framework to include graded intensity or richer emotion taxonomies such as sadness, anxiety, joy, or frustration.

6.4 Final Reflections

This study reinforces the vital role sentiment analysis plays in understanding how users emotionally experience digital mental health platforms. As these tools grow more widespread, especially in underserved areas, it becomes essential to ensure that they not only function technically but also connect empathetically and ethically. The integration of machine learning and transformer models with aspect-based analysis presents a promising future where mental health support is scalable, responsive, and emotionally intelligent.

REFERENCES

1. WHO. (2022). World mental health report: Transforming mental health for all. https://www.who.int/publications/i/item/9789240049338

2. Chisholm, D., Sweeny, K., Sheehan, P., Rasmussen, B., Smit, F., Cuijpers, P., & Saxena, S. (2016). Scaling-up treatment of depression and anxiety: A global return on investment analysis. The Lancet Psychiatry, 3(5), 415–424. https://doi.org/10.1016/S2215-0366(16)30024-4

3. Mutiso, V. N., Musyimi, C. W., Nayak, S. S., Tele, A., Musau, A. M., Rebello, T. J., Pike, K. M., & Ndetei, D. M. (2021). Formative research on the use of the WHO-AIMS tool to assess mental health systems in Makueni County, Kenya. International Journal of Mental Health Systems, 15(1), 1–12. https://doi.org/10.1186/s13033-021-00481-z

4. Ministry of Health, Kenya. (2021). Kenya Mental Health Action Plan 2021–2025. Ministry of Health. https://mental.health.go.ke/download/kenya-mental-health-action-plan-2021-2025/

5. Shan, Y., Ji, M., Xie, W., Lam, K.-Y., & Chow, C.-Y. (2022). Public Trust in Artificial Intelligence Applications in Mental Health Care: Topic Modeling Analysis. JMIR Human Factors, 9(4), e38799. https://doi.org/10.2196/38799

6. Bharadiya, J. P. (2022). DRIVING BUSINESS GROWTH WITH ARTIFICIAL INTELLIGENCE AND BUSINESS INTELLIGENCE. 6(4).

7. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5

8. Podder, I., Fischl, T., & Bub, U. (2023). Artificial Intelligence Applications for MEMS-Based Sensors and Manufacturing Process Optimization. Telecom, 4(1), 165–197. https://doi.org/10.3390/telecom4010011

9. Risdin, F., Mondal, P. K., & Hassan, K. M. (n.d.). Convolutional Neural Networks (CNN) for Detecting Fruit Information Using Machine Learning Techniques.

10. Friedrich, S., Groß, S., König, I. R., Engelhardt, S., Bahls, M., Heinz, J., Huber, C., Kaderali, L., Kelm, M., Leha, A., Rühl, J., Schaller, J., Scherer, C., Vollmer, M., Seidler, T., & Friede, T. (2021). Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: A systematic review with recommendations. European Heart Journal - Digital Health, 2(3), 424–436. https://doi.org/10.1093/ehjdh/ztab054

11. Mijwil, Maad M.; Aggarwal, Karan; Sonia, Sonia; Al-Mistarehi, Abdel Hameed; Alomari, Safwan; Gök, Murat; Zein Alaabdin, Anas M.; and Abdulrhman, Safaa H. (2022) "Has the Future Started? The Current Growth of ArtificialIntelligence, Machine Learning, and Deep Learning," Iraqi Journal

for Computer Science and Mathematics: Vol. 3: Iss. 1, Article 13. DOI: https://doi.org/10.52866/ijcsm.2022.01.01.013

12. Saravanan, R., & Sujatha, P. (2018). A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 945–949. https://doi.org/10.1109/ICCONS.2018.8663155

13. Mahesh, B. (2018). Machine Learning Algorithms—A Review. 9(1).

14. Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: Global trends, research directions, and regulatory standpoints. Data Science and Management, 4, 19–29. https://doi.org/10.1016/j.dsm.2021.12.002

15. Omuya, E. O., Okeyo, G., & Kimwele, M. (2023). Sentiment analysis on social media tweets using dimensionality reduction and natural language processing. Engineering Reports, 5(3), e12579. https://doi.org/10.1002/eng2.12579

16. El-Ansari, A., & Beni-Hssane, A. (2023). Sentiment Analysis for Personalized Chatbots in E-Commerce Applications. Wireless Personal Communications, 129(3), 1623–1644. https://doi.org/10.1007/s11277-023-10199-5

17. Haque, M R., & Rubya, S. (2023, May 22). An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. JMIR Publications, 11, e44838-e44838. https://doi.org/10.2196/44838

18. Abd-Alrazaq, A., Rababeh, A., Alajlani, M., Bewick, B M., & Househ, M. (2020, July 13). Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. JMIR Publications, 22(7), e16021-e16021. https://doi.org/10.2196/16021

19. GSMA. (2023). The mobile economy Sub-Saharan Africa 2023. https://www.gsma.com/mobileeconomy/sub-saharan-africa/

20. Følstad, A., & Brandtzæg, P B. (2020, April 11). Users' experiences with chatbots: findings from a questionnaire study. Springer Science+Business Media, 5(1). https://doi.org/10.1007/s41233-020-00033-2

21. Gkinko, L., & Elbanna, A. (2022). Hope, tolerance and empathy: employees' emotions when using an AI-enabled chatbot in a digitalised workplace. Information Technology & People, 35(6), 1714-1743.

22. Lu, Q., Sun, X., Long, Y., Gao, Z., Feng, J., & Sun, T. (2023). Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. IEEE Transactions on Neural Networks and Learning Systems, 35(11), 15092–15112. https://doi.org/10.1109/TNNLS.2023.3294810

23. Mohammad, S. M. (2016). A Practical Guide to Sentiment Annotation: Challenges and Solutions. In Proceedings of the 7th Workshop on

Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 174–179). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-0429

24. Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). SentiBench: A benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5(1), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

25. Taboada, M., Brooke, J., Tofilovski, M., Voll, K. V., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267–307.

26. Yadollahi, A., Shahraki, A. G., & Zaïane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys, 50(2), Article 25, 1–33. https://doi.org/10.1145/3057270

27. Dai, Y., Liu, J., Zhang, J., Fu, H., & Xu, Z. (2021). Unsupervised sentiment analysis by transferring multi-source knowledge. arXiv. https://doi.org/10.48550/arXiv.2105.11902

28. Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(6), 292–303. https://doi.org/10.1002/widm.1169

29. Shin, B., Lee, T., & Choi, J. D. (2016). Lexicon-integrated CNN models with attention for sentiment analysis. arXiv. https://arxiv.org/abs/1610.06272

30. Ahmed, A., Aziz, S., Khalifa, M., Shah, U., Hassan, A., Abd-Alrazaq, A., & Househ, M. (2022). Thematic analysis on user reviews for depression and anxiety chatbot apps: Machine learning approach. JMIR Formative Research, 6(3), e27654. https://doi.org/10.2196/27654

31. Khan, Z. A., Xia, Y., Aurangzeb, K., Khaliq, F., Alam, M., Khan, J. A., & Anwar, M. S. (2024a). Emotion detection from handwriting and drawing samples using an attention-based transformer model. PeerJ Computer Science, 10, e1887. https://doi.org/10.7717/peerj-cs.1887

32. Hernández Farías, D. I., & Rosso, P. (2016). Irony, sarcasm and sentiment analysis. In F. A. Pozzi, E. Fersini, E. Messina, & B. Liu (Eds.), Sentiment Analysis in Social Networks (pp. 113–128). Morgan Kaufmann.

33. Karoo, K., & Chitte, V. (2023). Ethical considerations in sentiment analysis: Navigating the complex landscape. International Research Journal of Modernization in Engineering Technology and Science, 5(11), 239–248. https://doi.org/10.56726/IRJMETS46811

34. Naslund, J. A., Aschbrenner, K. A., Araya, R., Marsch, L. A., Unützer, J., Patel, V., & Bartels, S. J. (2017). Digital technology for treating and preventing mental disorders in low-income and middle-income countries: A narrative review of the literature. The Lancet Psychiatry, 4(6), 486–500. https://doi.org/10.1016/S2215-0366(17)30096-2

35. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

36. Chancellor, S., Baumer, E. P. S., & De Choudhury, M. (2019). Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–32. https://doi.org/10.1145/3359249

37. Miah, M. S. U., Sulaiman, J., Sarwar, T. B., Zamli, K. Z., & Jose, R. (2021). Study of Keyword Extraction Techniques for Electric Double-Layer Capacitor Domain Using Text Similarity Indexes: An Experimental Analysis. Complexity, 2021, 1–12. https://doi.org/10.1155/2021/8192320

38. Ophir, Y., & Jamieson, K. H. (2020). The Effects of Zika Virus Risk Coverage on Familiarity, Knowledge and Behavior in the U.S. – A Time Series Analysis Combining Content Analysis and a Nationally Representative Survey. Health Communication, 35(1), 35–45. https://doi.org/10.1080/10410236.2018.1536958

39. Scarborough, P., Adhikari, V., Harrington, R. A., Elhussein, A., Briggs, A., Rayner, M., Adams, J., Cummins, S., Penney, T., & White, M. (2020). Impact of the announcement and implementation of the UK Soft Drinks Industry Levy on sugar content, price, product size and number of available soft drinks in the UK, 2015-19: A controlled interrupted time series analysis. PLOS Medicine, 17(2), e1003025. https://doi.org/10.1371/journal.pmed.1003025

40. Ravi, K., & Ravi, V. (2019). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems, 89, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

41. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

42. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19. https://doi.org/10.2196/mental.7785

43. Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. JMIR Mental Health, 5(4), e64. https://doi.org/10.2196/mental.9782

44. Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. JMIR Mental Health, 5(4), e64. https://doi.org/10.2196/mental.9782

45. Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation. JMIR mHealth and uHealth, 6(11), e12106. https://doi.org/10.2196/12106

46. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. Canadian Journal of Psychiatry, 64(7), 456–464. https://doi.org/10.1177/0706743719828977

47. Guntuku, S. C., Sherman, G., Stokes, D. C., Agarwal, A. K., Seltzer, E., Merchant, R. M., & Ungar, L. H. (2020). Tracking mental health and symptom mentions on Twitter during COVID 19. Journal of General Internal Medicine, 35(9), 2798–2800. https://doi.org/10.1007/s11606-020-05988-8

48. Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research, 63, 743–788. https://doi.org/10.1613/jair.1.11259

49. Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. Journal of Medical Internet Research, 15(11), e239. https://doi.org/10.2196/jmir.2721

50. Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. Artificial Intelligence Review, 53(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

51. Villanueva Miranda, I., Zhang, L., Khan, S., & Lee, J. A. (2025). Sentiment analysis in public health: A systematic review of the current state, challenges, and future directions. Frontiers in Public Health, 13, 1609749. https://doi.org/10.3389/fpubh.2025.1609749

52. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30–38). https://aclanthology.org/W11-0705/

53. AlSagri, H. S., & Ykhlef, M. (2020). Machine learning–based approach for depression detection in Twitter using content and activity features. IEICE Transactions on Information and Systems, E103-D(12), 1825–1832. https://doi.org/10.1587/transinf.2020EDP7023

54. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

55. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

56. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. https://doi.org/10.48550/arXiv.1904.03323

57. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805

58. Wagay, F. A., & Jahiruddin, J. (2025). Classification of mental illnesses from Reddit posts using Sentence BERT embeddings and neural networks. Procedia Computer Science, 258, 1669–1676. https://doi.org/10.1016/j.procs.2025.04.398

59. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

60. Haque, R., & Rubya, S. (2022). "For an app supposed to make its users feel better, it sure is a joke" – an analysis of user reviews of mobile mental health applications. arXiv. https://arxiv.org/abs/2209.07796

61. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In Machine Learning for Healthcare Conference (pp. 359–380). PMLR.

62. Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 7184–7190). European Language Resources Association. https://aclanthology.org/2022.lrec-1.778/

63. Alkhnbashi, O. S., Mohammad, R., & Hammoudeh, M. (2024). Aspect-based sentiment analysis of patient feedback using large language models. Big Data and Cognitive Computing, 8(12), 167. https://doi.org/10.3390/bdcc8120167

64. Aryanti, F. A. D., Luthfiarta, A., & Soeroso, D. A. I. (2025). Aspect-based sentiment analysis with LDA and IndoBERT algorithm on mental health app: Riliv. Journal of Applied Informatics and Computing, 9(2), 361–375. https://doi.org/10.30871/jaic.v9i2.8958

65. Hua, Y. C., Denny, P., Wicker, J., & Taskova, K. (2024). A systematic review of aspect based sentiment analysis: Domains, methods, and trends. Artificial Intelligence Review, 57, 296. https://doi.org/10.1007/s10462-024-10906-z

66. Wu, Z., & Ong, D. C. (2020). Context-guided BERT for targeted aspect-based sentiment analysis. arXiv preprint arXiv:2010.07523. https://arxiv.org/abs/2010.07523

67. Xu, H., Shu, L., Yu, P. S., & Liu, B. (2020). Understanding pre-trained BERT for aspect-based sentiment analysis. arXiv preprint arXiv:2011.00169. https://arxiv.org/abs/2011.00169

68. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (pp. 19–30). https://doi.org/10.18653/v1/S16-1002

69. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2019). Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 207–212). https://doi.org/10.18653/v1/P16-2034

70. Stawarz, K., Preist, C., & Coyle, D. (2019). Use of smartphone apps, social media, and web-based resources to support mental health and well-being: Online survey. JMIR Mental Health, 6(7), e12546. https://doi.org/10.2196/12546

71. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

72. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4040–4054. https://doi.org/10.18653/v1/2020.acl-main.372

73. Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. NPJ Digital Medicine, 3, 43. https://doi.org/10.1038/s41746-020-0233-7

74. Losada, D. E., Crestani, F., & Parapar, J. (2017). Overview of eRisk: Early risk detection on the internet. In Proceedings of the 8th International Conference of the CLEF Initiative (pp. 346–360). CEUR-WS.org.

75. Smith, J. A., Patel, R. K., & Chen, L. (2024). Estimation of minimal data set sizes for machine learning in digital mental health interventions. NPJ Digital Medicine, 7, Article 36. https://doi.org/10.1038/s41746-024-01360-w

76. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

77. Choudhury, M. D., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 3267–3276. https://doi.org/10.1145/2470654.2466447

78. Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). MentalBERT: Publicly available pretrained language models for mental healthcare. arXiv preprint arXiv:2110.15621. https://arxiv.org/abs/2110.15621

79. Mezzi, R., Yahyaoui, A., Krir, M. W., Boulila, W., & Koubaa, A. (2022). Mental health intent recognition for Arabic-speaking patients using the Mini International Neuropsychiatric Interview (MINI) and BERT model. Sensors, 22(3), 846. https://doi.org/10.3390/s22030846

80. Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit social media forum. IEEE Access, 7, 44883–44893. https://doi.org/10.1109/ACCESS.2019.2909180

81. Jake-Schoffman, D. E., Silfee, V. J., Waring, M. E., Boudreaux, E. D., Sadasivam, R. S., Mullen, S. P., Carey, J. L., Hayes, R. B., Ding, E. Y., Bennett, G. G., & Pagoto, S. L. (2017). Methods for evaluating the content, usability, and efficacy of commercial mobile health apps. JMIR mHealth and uHealth, 5(12), e190. https://doi.org/10.2196/mhealth.8758

82. Heedzy. (2016). Heedzy web scraping tool documentation. Retrieved from http://heedzy.com/documentation

83. Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2024). Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations. arXiv. https://doi.org/10.48550/arXiv.2410.23432

84. Patel, R., & Passi, K. (2020). Sentiment analysis on twitter data of world cup soccer tournament using machine learning. IoT, 1(2), 14.

85. Gebauer, J., Tang, Y., & Baimai, C. (2008). User requirements of mobile technology: results from a content analysis of user reviews. Information Systems and e-Business Management, 6, 361-384.

86. McIlroy, S., Ali, N., Khalid, H., & E. Hassan, A. (2016). Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. Empirical Software Engineering, 21, 1067-1106.

87. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225.

88. Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Multi-faceted sentiment analysis. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2009), 50–58.

89. Bounabi, M., El Moutaouakil, K., & Satori, K. (2019, October). Text classification using fuzzy TF-IDF and machine learning models. In Proceedings of the 4th international conference on big data and Internet of Things (pp. 1-6).

90. Hu, K., Wu, H., Qi, K., Yu, J., Yang, S., Yu, T., ... & Liu, B. (2018). A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. Scientometrics, 114, 1031-1068.

91. Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. Expert systems with applications, 38(3), 2758-2765.

92. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

93. Santhosh Baboo, S., & Amirthapriya, M. (2022). Comparison of machine learning techniques on Twitter emotions classification. SN Computer Science, 3(1), 35.

94. Eriksson, T. (2013). Automatic web page categorizationusing text classication methods.

95. Phillips, J., Cripps, E., Lau, J. W., & Hodkiewicz, M. R. (2015). Classifying machinery condition using oil samples and binary logistic regression. Mechanical Systems and Signal Processing, 60, 316-325.

96. Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

97. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

98. Association of Internet Researchers. (2019). Ethical Guidelines for Internet Research (IRE 3.0). AoIR. Retrieved from https://aoir.org/reports/ethics3.pdf

99. Thompson, R., & Chen, L. (2020). Limitations of star ratings in sentiment analysis: A case study. User Experience Research Quarterly, 12(3), 201–215.

100. Gupta, R. (2019). Data augmentation for low resource sentiment analysis using generative adversarial networks. arXiv. https://doi.org/10.48550/arXiv.1902.06818

101. Gao, X., & Liu, Y. (2019). Towards a cross-platform evaluation framework for app success metrics. Information Systems Frontiers, 21(4), 865–880.

102. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186. https://doi.org/10.18653/v1/N19-1423

103. Elamin, M. (2020). Sentiment analysis challenges in text classification: A comprehensive review. International Journal of Computer Science and Information Technology, 12(3), 45–60.

104. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

105. Liu, Y., Smith, T., & Zhang, H. (2022). Advancements in transformer-based sentiment analysis for healthcare applications. Journal of Healthcare Informatics Research, 6(2), 123–145.

106. Zhou, Q., Zhang, Y., & Li, Z. (2020). Ensemble methods for emotion detection in social media. IEEE Transactions on Affective Computing, 11(3), 456–467.