NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

“Translational Engineering in Health and Medicine”

COMPARATIVE ANALYSIS FOR MENTAL HEALTH PREDICTION TASKS BASED ON SOCIAL MEDIA POSTS

Postgraduate Diploma Thesis

of

Postgraduate student: **Spiliotis Theodoros A.**

Supervisor : Nikita Konstantina

Professor, NTUA

Athens, June 2024

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

"Translational Engineering in Health and Medicine"

COMPARATIVE ANALYSIS FOR MENTAL HEALTH PREDICTION TASKS BASED ON SOCIAL MEDIA POSTS

Postgraduate Diploma Thesis

of

Postgraduate student: **Spiliotis Theodoros A.**

Supervisor : Nikita Konstantina
Professor, NTUA

The postgraduate diploma thesis has been approved by the examination committee on *(exam day: 8 July 2024)*

(Signature)	*(Signature)*	*(Signature)*
…….	……	…….
Nikita K.	Stamou G.	Voulodimos A.
Prodessor,NTUA	Professor, NTUA	Assistant Professor, NTUA

Athens, June 2024

…….

Theodoros A. Spiliotis

Graduate of the Interdisciplinary Postgraduate Programme,

“Translational Engineering in Health and Medicine”,

Master of Science,

School of Electrical and Computer Engineering,

National Technical University of Athens

Abstract

In this thesis, titled "Comparative Analysis for Mental Health Prediction Tasks Based on Social Media Posts," the challenge of detecting depression and suicidal ideation through natural language processing (NLP) on social media platforms is addressed. Depression, a prevalent and debilitating mental health disorder characterized by persistent feelings of sadness, hopelessness, and loss of interest, affects millions globally. The study aims to enhance early detection and intervention strategies by leveraging advanced machine learning models. Social media's widespread use presents a unique opportunity to analyze user-generated content for mental health insights, providing a non-intrusive method to monitor and support individuals at risk. Social media platforms like Twitter and Reddit offer rich, real-time data sources where individuals often express their emotions and mental states, making them valuable for detecting signs of mental health issues.

The research utilizes datasets from these platforms, focusing on depressive and suicidal content. Data preprocessing steps include tokenization, lemmatization, and feature extraction using techniques such as TF-IDF and word embeddings. Various machine learning models, including Logistic Regression, Random Forest, and state-of-the-art transformer-based models like BERT and DistilBERT, are fine-tuned for the task. Comparative analysis between these models highlights their respective strengths and weaknesses in detecting depressive and suicidal language.

The evaluation reveals that transformer-based models, particularly DistilBERT, significantly outperform traditional machine learning methods in accuracy, precision, recall, and F1-score. For instance, DistilBERT achieved an F1-score of 0.99 on the Depression Twitter dataset, highlighting its capability to discern depressive content with high precision and recall. Comparatively, the Random Forest classifier also showed strong performance but was slightly outpaced by the transformer models. This comparative analysis provides valuable insights into the effectiveness of different machine learning approaches for mental health prediction.

The thesis underscores the need for further research into the integration of multimodal data, including textual and non-textual inputs such as images and user interaction patterns. Future studies should focus on the ethical considerations of using social media data, ensuring user privacy and consent. Additionally, continuous model adaptation and fine-tuning to evolving linguistic trends and emerging data sources will be crucial for maintaining model accuracy and relevance. Addressing these ethical and technical challenges is essential for the responsible deployment of these technologies in real-world mental health support systems.

In conclusion, this research demonstrates the potential of NLP and machine learning in monitoring and supporting mental health through social media analysis. By harnessing the vast and diverse data available on social media platforms, we can develop proactive measures to identify and assist individuals at risk of depression and suicidal ideation. The findings advocate for the integration of these technologies into mental health services to provide timely and accurate interventions, ultimately contributing to improved mental health outcomes. This approach represents a significant step forward in the ongoing effort to leverage digital technologies for better mental health care.

Keywords: Depression detection, suicidal ideation, natural language processing, social media analysis, machine learning, mental health prediction, transformer models, comparative analysis, ethical considerations, multimodal data.

Introduction

This thesis was conducted at the Laboratory of Biomedical Simulations and Imaging Technology (BIOSIM) of the National Technical University of Athens and explores the intricate relationship between mental health and social media, focusing on how digital interactions influence psychological well-being and the potential for leveraging technology in mental health monitoring and intervention. The content is organized into five comprehensive chapters, each addressing different aspects of this relationship and presenting both theoretical and empirical insights.

Chapter 1 lays the foundation by examining the broad impact of social media on mental health. It begins with an in-depth discussion on depression and suicide, highlighting how these mental health issues are exacerbated by social media use. This chapter also delves into the psychological, positive, and negative emotional influences of social media usage. Further, it investigates the correlation between social media usage patterns and depression, as well as how social media can exacerbate depressive symptoms. Finally, it addresses the connection between social media patterns and suicide, providing a comprehensive overview of the mental health challenges posed by digital platforms.

Chapter 2 focuses on advanced technological methods for monitoring mental health. It starts with an exploration of machine learning techniques, including supervised learning, neural networks, and evaluation metrics. The chapter then transitions to the relevance of Natural Language Processing (NLP) in mental health monitoring, detailing specific techniques and metrics used for detecting depressive content. It also covers the application of Large Language Models (LLMs) in mental health prediction, discussing their potential and challenges. Additionally, the chapter addresses the ethical considerations and challenges in data collection, ensuring privacy and accuracy in mental health monitoring. It concludes with a discussion on the regulatory and compliance issues related to mental health data.

Chapter 3 provides a detailed overview of the research methodology employed in the thesis. It outlines the datasets used for analysis, including those related to depression and suicide from various social media platforms like Twitter and Reddit. The chapter describes the specific NLP methodologies applied to these datasets, such as BERT for depression detection and techniques for suicide prediction from Twitter. This methodological framework sets the stage for the empirical analysis conducted in subsequent chapters.

Chapter 4 presents the findings from the empirical analysis of the social media datasets. It begins with results from the analysis of Twitter data, including the performance of various models such as Naive Bayes with TF-IDF. The chapter continues with a comparative analysis of depression datasets from Twitter and Reddit, providing insights into the effectiveness of different approaches in detecting depressive and suicidal content. The results highlight key patterns and trends in social media data, emphasizing the potential of

The final chapter synthesizes the findings from the research and discusses their implications for mental health monitoring and intervention. It outlines the contributions of the thesis to the field of mental health research and suggests areas for future work. This chapter emphasizes the importance of continued innovation in leveraging technology for mental health support and highlights the need for ongoing ethical and regulatory considerations.

Closing , I would also like to add some acknowledgements.

I would like to extend my heartfelt gratitude to my supervising professor, Konstantina Nikita, for assigning me this interesting topic and for the trust she showed in allowing me to thoroughly engage with it.

Additionally, I wish to express my sincere thanks to Dr. Kostas Mitsis for the excellent collaboration during the preparation of this thesis and for the invaluable time he dedicated to assisting me. His valuable advice and significant contribution were crucial in achieving the final goal of this work.

I would also like to thank my friends, especially those I met through the university. Kostas, Kontessa, Thenia, Thanos, Marina, Antonis—meeting you was the most important gain from my years at university.

Moreover, I would like to specially thank Professor Leonidas Alexopoulos and Vasilis Papakonstantinou for their support in the Biodesign Innovative Process course and for the inspiration they provided. Through this postgraduate program, one of my dreams began to take shape: to engage in entrepreneurship in the healthcare sector, where I feel I can have a positive impact on the world.

Lastly, I express my profound gratitude to my family, my parents Antonis and Panagiota, my sisters Eleni and Martha, and my partner Eirini for their encouragement and support throughout my studies.

Contents Table

Figure's Table

Chapter 1: Mental Health and Social Media

In today's digital era, the relationship between mental health and social media has become critically important. Social media platforms, while providing unprecedented connectivity and avenues for self-expression, also present unique challenges and risks to mental health. These platforms have transformed the way individuals interact, share experiences, and perceive themselves, often amplifying both positive and negative aspects of human behavior. This chapter explores the complex relationship between mental health issues, such as depression and suicide, and the pervasive influence of social media. By examining current research and trends, we aim to understand how these platforms affect mental health, both positively and negatively, and identify potential strategies for mitigation and support. Continuous exposure to idealized images and lifestyles can lead to feelings of inadequacy, anxiety, and depression, while the anonymity and reach of social media can sometimes facilitate harmful behaviors such as cyberbullying. Conversely, social media also offers opportunities for building supportive communities and accessing mental health resources. The chapter begins with an in-depth look at depression, a prevalent and debilitating mental health disorder, examining its symptoms, causes, and the impact of social media. It then progresses to discuss suicide, another critical mental health issue, and finally, broader implications of social media on mental well-being, including both the risks and the potential benefits, providing a comprehensive overview of how digital interactions shape our psychological health in contemporary society[1].

1.1. Mental Health

Mental health encompasses a wide range of emotional, psychological, and social well-being aspects that significantly impact how individuals think, feel, and act. It is crucial for determining how we handle stress, relate to others, and make choices. Mental health issues, such as depression and suicide, pose serious challenges to individuals and societies worldwide. According to Zalsman [2], mental health disorders like depression are prevalent, affecting approximately 5% of adults globally, and are becoming increasingly common. These disorders can lead to severe consequences, including diminished functioning, well-being, and in severe cases, self-harm or suicide. Addressing mental health issues requires early identification, timely intervention, and comprehensive strategies that encompass traditional clinical methods and modern technological advancements. In this thesis, we will specifically explore the complex issues of depression and suicide, examining how they are influenced by social media and how emerging technologies can aid in their detection and prevention[2].

1.1.1 Depression

Depression, also known as depressive disorder, is a widespread mental illness with severe implications for individuals' functioning and well-being, including the potential for self-harm. It is particularly concerning in adolescence and can persist into adulthood, affecting approximately 5% of adults globally, with a higher prevalence in middle-aged individuals. Moreover, depression rates have been steadily increasing worldwide from 2005 to 2022 [3].

Early identification and intervention are crucial for managing depression and reducing its severity and adverse effects on individuals' lives. While traditional methods rely on clinical procedures and surveys to diagnose depression, there is a growing need for automatic depression detection systems to streamline the process and improve outcomes. The rise of the internet and social media platforms has provided new avenues for individuals to express their emotions and thoughts, with text emerging as a primary medium for sharing among those experiencing depression [3].

Advancements in artificial intelligence and deep learning techniques have paved the way for creating more efficient depression detection systems. Natural language processing (NLP) techniques, including word representations, play a crucial role in identifying linguistic patterns associated with depression in text data. As such, there is a significant demand for enhancing embedding methods and learning techniques to improve depression detection accuracy from textual content[3].

According to Kabir [4] depression in low-income often goes unrecognized and lacks adequate mental health services, leaving many individuals without the support they need. Access to these services is severely limited, which further exacerbates the problem. The stigma associated with mental health disorders also discourages many from seeking help. However, research indicates that early and accurate treatment significantly increases the likelihood of a full recovery.

In this context, Bhoge[5] highlights the critical role of predictive analytics in improving the early detection of depression. By developing sophisticated algorithms and machine learning techniques, researchers can analyze behavioral and social network patterns to predict depressive episodes. This advancement allows for more precise and timely interventions, potentially setting up more effective treatment plans.

Depression is a complex and multifaceted issue that affects billions of people worldwide. This research underscores the urgency of addressing depression by evaluating its symptoms, consequences, and detection methods. It emphasizes the need to enhance our understanding and develop new approaches to treatment and prevention.

One promising area of development is the use of natural language processing (NLP) to scan social media feeds for signs of depression. While this is just one piece of the puzzle, it represents a significant step forward in the ongoing battle against this debilitating condition. By leveraging NLP, we can potentially identify individuals at risk and provide them with the necessary resources and support to manage their mental health effectively.

1.1.2 Suicide

Suicide is a critical public health issue that requires comprehensive and multifaceted approaches for prevention. The complexity of the factors leading to suicide necessitates various strategies that target different levels of intervention. Here, we examine evidence from multiple research studies to understand the effectiveness of different suicide prevention strategies and identify best practices for reducing suicide rates.

Research consistently highlights the effectiveness of restricting access to means of suicide as a crucial strategy for prevention. According to a systematic review by Zalsman [2], measures such as controlling access to analgesics and implementing barriers at known suicide hotspots significantly reduce suicide rates. The review found a 43% reduction in suicides related to analgesic overdose and an 86% reduction in suicides at jumping hotspots [2]. These findings underscore the importance of environmental modifications in suicide prevention.

School-based programs have shown promising results in reducing suicidal ideation and attempts among adolescents. Zalsman [2] noted that such programs could decrease the odds of suicide attempts (OR 0.45) and suicidal ideation (OR 0.50). These programs typically involve educating students about mental health, recognizing warning signs, and promoting help-seeking behaviors. The effectiveness of these interventions suggests that early education and awareness can play a significant role in mitigating risk factors associated with suicide [2].

Pharmacological treatments, particularly the use of lithium and clozapine, have been substantiated as effective in reducing suicide rates among individuals with severe mental disorders. A meta-analysis highlighted by Riblet [6] showed that the World Health Organization's brief intervention and contact (BIC) significantly lowered the odds of suicide. Although lithium and clozapine have demonstrated efficacy, their use must be carefully managed to ensure safety and adherence [6].

Training general practitioners (GPs) to recognize and treat depression and suicidality is another critical strategy. Van der Feltz-Cornelis [7] identified that enhancing the skills of GPs in diagnosing and managing mental health issues can significantly impact suicide prevention. Improved accessibility of care and the ability to identify at-risk individuals early are essential components of this approach [7].

Multilevel interventions that combine various strategies are increasingly recognized for their potential synergistic effects. Hegerl [7] discusses the European Alliance Against Depression (EAAD), which implements a four-level approach including public awareness campaigns, training for healthcare providers, support for high-risk groups, and media engagement. Such comprehensive strategies are crucial for addressing the multifaceted nature of suicide and have shown effectiveness in different cultural contexts [7].

Enhancing patients' social networks and reducing the stigma associated with mental health can also significantly contribute to suicide prevention. Eagles [8] found that social support from friends and family, alongside professional psychiatric services, was considered equally helpful by individuals experiencing suicidal thoughts. Efforts to strengthen social connections and reduce isolation are therefore critical components of a comprehensive suicide prevention strategy [8].

In summary, effective suicide prevention requires a combination of strategies that include restricting access to lethal means, school-based awareness programs, pharmacological treatments, training for healthcare providers, multilevel community-based interventions, and strengthening social support networks. These strategies, when implemented together, can create a robust framework for reducing suicide rates and addressing the underlying factors contributing to suicidal behavior. Continued research and evaluation are necessary to optimize these interventions and ensure they are effectively tailored to different populations and settings.

1.2.Social Media Affects Emotions

Through its integration into the contemporary lifestyle, social media has resulted in a plethora of changes where the world of being connected to others and expressing oneself has been completely altered. The spread of currently accessible platforms leads to a manifestation of emotive conditions of users that may be displayed unusually or oppositely.

1.2.1. Psychological Effects of Social Media Usage

Social media platforms significantly influence users' mental health, both positively and negatively. On the positive side, social media provides powerful tools for social connection, allowing users to maintain relationships and receive support during difficult times, which is particularly beneficial for those with mobility or accessibility issues. For example, platforms enable users to stay connected with friends and family despite geographical distances, reducing feelings of loneliness and providing a support network [9].

Chancellor and De Choudhury [10] have made a comprehensive research on the mental health implications of social media, stating that these platforms can immensely influence a person's mental health. Social media can be a powerful tool for social connection, allowing users to connect with their friends despite the miles between them, and access a support network during difficult times. This is particularly important for those with mobility or accessibility problems, as social media becomes a medium to interact with their social group members, reducing loneliness.

However, the influence of social media is not uniformly positive. Just as online services offer encouragement, they can also lead to greater jealousy and inferiority through mechanisms like social comparison. Frequently, the reality doesn't reach the expected level of a person's future because of creating an image of other people's lives that looks like an

"idealized representation" online, which makes users feel envy and inadequacy, and lowers their self-esteem. Similarly, the relentless supply of communications can result in an oversupply of information and demands from social life, which intensify stress and contribute to the emergence of depressive symptoms [11].

Excessive social media usage can also result in psychological distress. Studies have found that high amounts of personal social media usage are associated with lower task performance, increased technostress, and reduced happiness levels[12]. Furthermore, constant exposure to an overwhelming stream of information and social demands can intensify stress and negatively impact mental health.

1.2.2. Positive Emotional Influences of Social Media

Social media platforms play a crucial role in providing emotional support and fostering community development. They serve as a bridge that connects people, allowing them to maintain relationships that might otherwise be constrained by distance. This is particularly beneficial in times when many friends and relatives live in different towns or continents. Sharing moments through photos, videos, or status updates helps communities stay emotionally connected despite physical separation[13].

Social media provides substantial emotional support by allowing users to form peer support groups. These groups are especially important for individuals dealing with chronic illnesses, mental health issues, or personal problems. These online communities offer a space where people can share personal experiences, offer advice, and reduce the stigma surrounding various conditions. Campaigns of awareness and personal stories posted on social media can educate the public and build more compassionate community values[9] [10].

Furthermore, social media enables immediate feedback and empathy, which can be crucial during times of distress. The ability to connect easily with one's network for instant advice or comfort can make users feel more supported. Platforms also promote positive and motivational content that can improve users' moods and broaden their horizons [13].

Social media platforms facilitate community development by uniting individuals with common interests and goals. For instance, they enable the formation of support groups that provide emotional and informational support, enhancing users' sense of belonging and satisfaction). These online communities can significantly alleviate the emotional stress associated with conditions like autism by providing a platform for caregivers to share, gather, and exchange information. Additionally, the ability to maintain and build relationships through social media contributes to the development of social capital. This is the social equivalent of financial capital, which fosters collaboration and cooperation within communities, making life easier for their members[14].

1.2.3. Negative Emotional Influences of Social Media

Social media's negative emotional influences are multifaceted and can deeply affect users' state of mind. In detail, Chancellor and De Choudhury [10] and Ahmed [13] introduce the negative effects of social media that deserve special consideration. A similar problem consists of comparing cultures that spread on social media networks, where users like to compare their ordinary lives with a set of representative lives presented by other peers. Such comparisons frequently lead to a decline in the level of self-esteem, the lament of unnecessary feelings of insufficiency, and permanent dissatisfaction with one’s life. An inherent problem in the social media content cycle implies that people essentially are exposed to the best and brightest moments in others' lives, and this little bit of information is enough to create an artificial and distorted perception of normal life challenges and achievements [10]. Together with that, there is the concept of the “fear of missing out” (FOMO) which also makes matters worse. As someone looks through their feed, full of nonstop everyday updates, they might feel

like they are being left out of all the fun and excitement. This could worsen the anxiety of users and they may resort to their devices to cope with it, which then could act as distractions and obstruction to their concentration and face-to-face social interactions [13]. Social media's frequent usage often depletes stress levels with the intensity of impact on mental health and the manner the individual interacts with their environment brought on by this.

Moreover, the spread of misinformation and disinformation on social media platforms contributes to many people having more anxiety and confusion. In the digital age we live in, false information can quickly circulate, creating an atmosphere of anxiety and panic that can be nearly impossible to reign in or neutralize. Social media algorithms generally tend to give precedence to the content that engages the users the most, sometimes it is comprised of sensational and controversial information that might not be correct [10]. These can result in an ill-informed community intimidating others. This act is against the principle of informing the public and free discussions. Moreover, the effect of social media on sleep must not be undervalued as well. The blue light from screens is one of the factors interfering with the natural rest-activity cycle resulting in sleeping problems such as insomnia. Poor sleep has rather been shown to accompany many mental conditions such as depression and anxiety. Individuals who use social media at bedtime may have their sleep quality interfered with at levels non-conducive to their nervous and physical well-being [13]. These issues thereby ignite and catalyze a call for digital detoxing and for better management of online time. Medical professionals observe that unplugging from digital devices for some time accounts for a reduction of stress. Therefore, social media platforms, have their share of the responsibility and should curb cyberbullying, refine the information people share, and design algorithms that together with the other users create positive interactions and true connections [10] [13]. In summary, when it is used properly social media is very helpful and can give people so much but the effect on mental health poses a big challenge and needs close monitoring.

1.3. Social Media and Depression

The intersection of social media usage and mental health problems, namely depression, has become the focal point of many studies. Eichstaedt[15], Calvo[16], Kim[17], and Yam[18] research shows that social media patterns are linked with depression and the symptoms might be made worse by this.

1.3.1. Correlation between Social Media Usage Patterns and Depression.

Eichstaedt[15] attests to a significant relationship between the manner and frequency of social media consumption and the degree of depressive symptoms among users. With the sophisticated use of natural language processing (NLP) techniques Eichstaedt's research has employed, it was possible to identify the language patterns of posts that can predict feelings of depression. For example, words expressing negative emotions, self-focus, and isolation were more frequent in the posts of people who have depression than in a control group. Digital phenotyping could provide non-intrusive methods for remote monitoring of depression symptoms among groups of people. The method implemented by Calvo [16] was applying NLP in posts from different social media platforms. The conclusion was that long-term excessive use of social media increases the risk rate of developing depression. Their study underlined that the amount of information that a person passively absorbs, such as newsfeeds scrolled through without active engagement, was the highest reason for having elevated depression rates. This type of communication seems to boost the sense of loneliness and envy which are well-known triggers for depression. Generally, it seems that social media plays an important role in NLP research for preventing mental health illness. Based on Figure 1, Zhang[19] made a comprehensive study with 7536 bibliographic articles and we found that social media are the most used source of data for NLP applications. Especially Twitter and Reddit are the two most utilized social media platforms.

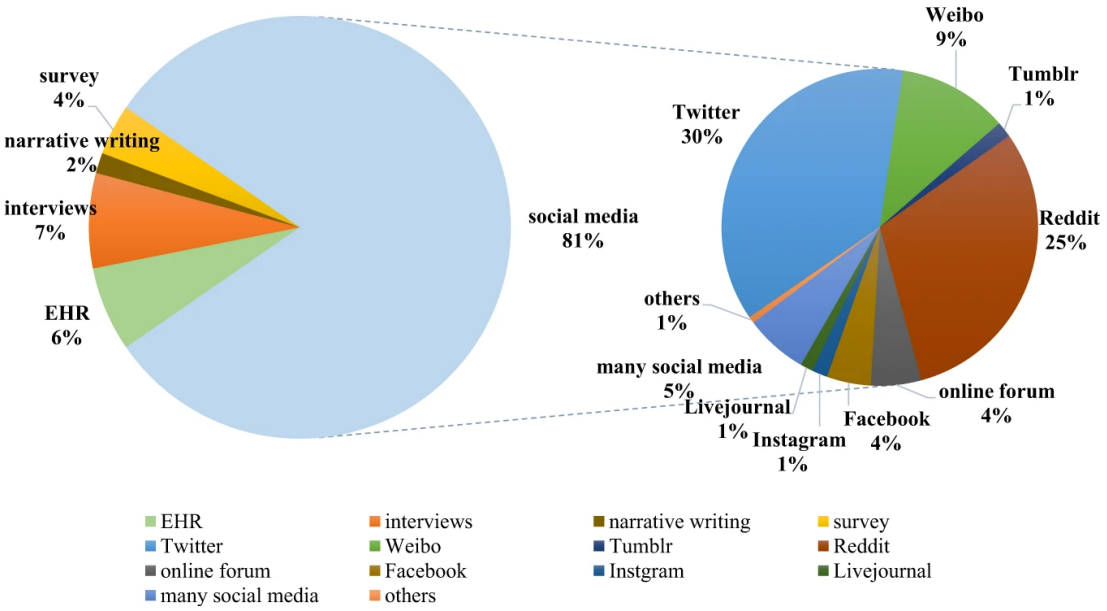

Figure 1:Distribution of different data sources[19]

1.3.2. How Social Media May Exacerbate Depressive Symptoms

Kim [17] furthers the discussion on the outcome of social media use by highlighting the nature in which these platforms can bring forth an increase in depressive symptoms. Their research considered the fact that people who engage in such social media image comparison feel less adequate or cheerless more often compared to those who do not. The purpose which people tend to use social media content, in most cases, is to showcase the best side of their lives which is just the imaginary version of reality, which can increase the worthiness of the problems, leading to greater self-deprecating thoughts in depressive users. Yam [18] goes one step further in exploring the relationship between depression and social media regarding its feedback loop. Such data gives support to the idea that depressed people may have a specific routine or pattern of social media use that makes their depression even more severe and hopeless. For instance, Facebook addiction can be an issue lending to internet community formation around depressing content and this exposure can bring a sense of belonging at the same time can reinforce and normalize negative thinking. On the other hand, instant but transitory dopamine released by likes and comments makes it possible for people to become addicted to them, a situation that hinders the creation of meaningful personal relationships, and other enjoyable activities. This will make the person fall into a state of loneliness and dissatisfaction.

The above studies have shown that emotional reactivity to social media is often associated with frequent access and the quick scrolling down of the content due to the design features that encourage such behavior. These elements tend to be more harmful to people who are already at risk for or suffering from depression illness. The uninterrupted upstream of notifications and updates can break attention, trigger stress, and impair sleep, thus aggravating signs like exuberance, fatigue, and mood swings. Because of the complexity and the that factors vary from one person to another, the link between social media use and depression is influenced by personality traits, the specific type of social media engagement, and the kind of content consumed. The fact that social media provides the most amazing ways of communicating and getting help leads to the fact that its misuse or overuse can only make mental illnesses worse, especially among vulnerable people. With comprehension growing concerning the way social media exploitation affects mental health, the use of strategies to

reduce the negative impacts of social media exploitation becomes increasingly crucial. This can be translated as adopting more responsible social media designs, advocating digital literacy that focuses on healthy use of such, incorporating mental health resources into these spaces, and providing support to people experiencing difficulties. Therefore, recognizing and dealing with the dual impact of social media on depression is paramount. By developing a balanced position on technology users, society will simplify exploiting the advantages and be able to guard against the risks, hence, having incorporated the digital presence, society will have better mental health strategies.

In the analysis of social media data to differentiate between individuals with and without depression, seven key features were highlighted that potentially exacerbate depressive symptoms through social media usage. Firstly, the level of education influences how openly individuals express symptoms and signs of depression, prompting a need for additional research to explore the correlation between educational attainment and depressive symptoms. Secondly, age and relationship status are associated with varying risks of depression; single individuals and those with partners tend to experience an increased risk with age, while married individuals typically exhibit a lower incidence of depression. Thirdly, a noticeable gender disparity exists, with a lower proportion of depression observed in males compared to females, which points to a significant relationship between gender and depression. Fourthly, the size of one's social network plays a crucial role; users with smaller social circles who post more depression-related markers are more likely to suffer from depression, indicating a tendency toward social isolation among individuals with depression. Fifthly, the frequency of participation in social activities shows a negative correlation with depression, where individuals without depression are more actively engaged in social activities. Sixthly, the extent of Facebook activity correlates with how long users engage with the platform and is linked to receiving more social invitations. Lastly, the usage of social features such as the 'like' function and showing interest in friends' activities is less common among individuals with depression, reflecting a reduced level of interaction and engagement in the social aspects of the platform. This comprehensive assessment reveals how different facets of social media usage can reflect or amplify depressive symptoms[20].

Finally, in another comprehensive research study[19] with 7536 papers, it seemed that depression and suicide are the most common mental illnesses explored (Figure 2) and that is the reason this research discovers those conditions.

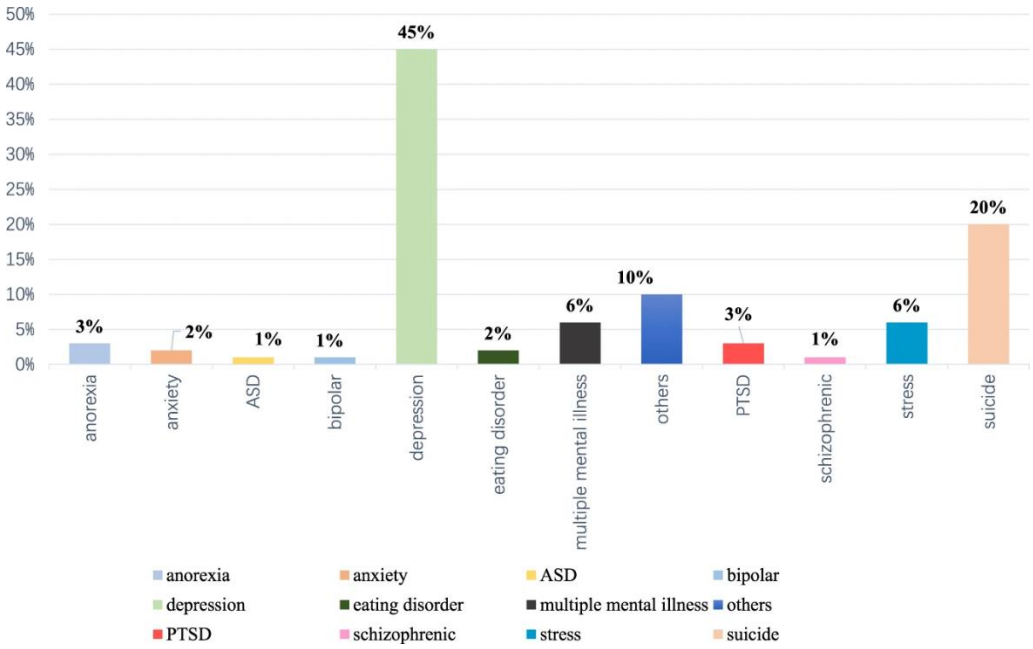

Figure 2:Proportions of various typers of mental illness[19]

1.4. Social Media and Suicide

We have already explored how social media can influence individuals toward depression, but the impact of these platforms extends beyond merely depressive symptoms. Social media generates vast amounts of data that not only reflect user behaviors and emotions but also provide critical insights predictive of suicidal tendencies. These platforms, as outlets for emotional expression, serve as valuable resources for detecting potential suicidal behavior.

The significance of social media data in suicide detection is paramount. Real-time insights into the behaviors and emotional states of individuals, which are traditionally challenging to capture, allow for early detection and intervention. Research by Coppersmith [21]and Aldhyani [22]has shown how analyzing linguistic patterns and engagement behaviors can effectively identify those at risk of suicide.

Identifying key patterns for suicide prediction involves analyzing several aspects of social media usage. Changes in the frequency and timing of posts, for example, can indicate underlying mental health states; an increase in activity during late hours might suggest insomnia or depression. Additionally, the sentiment expressed in posts and the presence of specific words or phrases related to despair or hopelessness are strong indicators of suicide risk. Machine learning models excel at processing vast amounts of textual data to identify these markers efficiently[21].

Another significant factor is the level of social interaction. A reduction in social interactions or changes in the type of interactions can indicate social withdrawal, a known risk factor for suicide[21].

The integration of advanced Natural Language Processing(NLP) and machine learning techniques has significantly improved the accuracy of predicting suicidal ideation through social media. NLP techniques are utilized to analyze the text for semantic and emotional content that may indicate distress or suicidal thoughts. Deep learning models, such as the CNN-BiLSTM model used by Aldhyani [22], are particularly effective in detecting complex patterns in textual data that signify suicidal ideation. Furthermore, hybrid models that incorporate multiple data types and sources—combining textual analysis with usage metrics like login frequency and interaction rates—enhance prediction accuracy [22].

In summary, the correlation between social media usage patterns and suicide underscores the potential of these digital platforms as tools for early detection and intervention in suicide prevention. By leveraging sophisticated machine learning and NLP techniques, researchers and practitioners can develop more effective tools to identify and support individuals at risk of suicide, ultimately contributing to better mental health outcomes
.

Chapter 2: Mental Health Monitoring

The rapid advancement of technology has revolutionized many fields, including mental health monitoring. As mental health issues such as depression and suicide continue to rise globally, there is an urgent need for innovative and effective methods to monitor and address these problems. This chapter delves into the state-of-the-art techniques and methodologies in mental health monitoring, exploring how emerging technologies can be leveraged to improve mental health outcomes.

Understanding these advanced methodologies is crucial for this thesis because it allows us to adopt and refine paradigms that are currently used in mental health monitoring. By examining these procedures, we can identify what is most important to test and validate, ensuring that our approach is both effective and scientifically robust.

Traditional methods of mental health monitoring, which primarily rely on clinical assessments and self-report surveys, are often insufficient due to stigma, accessibility issues, and infrequent clinical visits. In contrast, advanced technologies like machine learning and natural language processing (NLP) offer continuous and scalable solutions that can overcome these limitations.

This chapter covers machine learning techniques, including supervised learning and neural networks, and discusses evaluation metrics for these models. It then explores NLP's relevance to mental health monitoring, detailing techniques and metrics used to detect depressive content. The role of large language models (LLMs) in mental health prediction is examined, along with the use of chatbots for mental health support. Ethical considerations, challenges, and regulatory issues in mental health monitoring are also addressed. The chapter

concludes with a comparative analysis of NLP models and LLMs for depression prediction, highlighting their effectiveness and potential applications.

2.1. Machine Learning

Machine learning (ML) is a broad field applied to information technology, statistics, probability, artificial intelligence, psychology, neurobiology and many other disciplines. ML can solve problems simply by constructing a model that is a good representation of the database that reflects the problem. It is an advanced field of teaching computers to mimic the human brain and has brought to the field of statistics, fundamental computational theories of learning processes [23].

According to Zou [24] ML is defined as the field of study, which gives computers the ability to learn without being explicitly programmed for it. This area is used to teach machines how to manage data more efficiently. Through a range of algorithms, it seeks to interpret information that can be extracted from data and is not apparent simply through its projection. With the abundance and swelling of available datasets, the demand for ML is growing. Its purpose is to learn from data. In fact, there have been many studies on how humans will teach machines to learn on their own, without being explicitly programmed for it. Many mathematicians and programmers apply various approaches to find the solution to this problem, which requires the exploitation of huge databases[24].

ML algorithms are created to represent a human's ability to learn a task. These days, the development of new technologies in the field of Big Data has brought these algorithms in new forms, improved and renewed, and therefore their development is now about how they manage to process an increasing amount of data in less time, with little or no impact on results[23].

Based on the process followed by the algorithm, ML is divided into the following four ML techniques:

Supervised Learning, which uses knowledge from historical data, which has been mapped to tags/labels. Initially, this data is given as inputs to the system, along with the correct tags for the purpose of training the model. In this process, the model discovers correlations and patterns between the characteristics of the data and its correct labels, so that it can predict the label for unknown data that will be provided as future input. Essentially, the algorithm compares the obtained results with the actual expected ones to identify errors and thus adjust the model [25]. In the context of this thesis, we use this technique of ML to develop an emotion recognition model and therefore supervised learning is explained more extensively in the following sections.

Unsupervised Learning, which is used when the training features are not labeled. Through this technique, systems are developed that can infer a function that explains the hidden patterns present in unlabeled data. The system does not determine the correct result, but discovers from data specific matches and groups them [16].

Semi-supervised learning, which is a mix of supervised and unsupervised learning, as tagged and untagged data is utilized. It generally considers a smaller amount of labeled data and a larger amount of unlabeled data. These types of techniques can be adapted to achieve more accurate results. This technique is preferred in cases where the available tagged data needs other appropriate resources to be trained by them [16].

Reinforcement Learning is the technique that interacts with the environment with actions and identifies errors and rewards. Test error methods and delayed rewards are some of the common features of reinforcement learning. It enables the system and software programs to determine the ideal behavior of a particular content and increase its performance [16].

Lastly, Deep Learning, a subset of ML field, further enhances these techniques by employing neural networks that can automatically learn and extract complex patterns from large datasets. Unlike classical ML techniques that often rely on manual feature extraction,

deep learning models use multiple layers of neural networks to progressively refine data representations. This enables deep learning to excel in tasks such as image recognition, natural language processing, and predictive analytics. By building on the foundations laid by traditional ML methods, deep learning has revolutionized the ability of machines to process and understand vast amounts of data with unprecedented accuracy and efficiency.

2.1.1. Supervised Learning

The learning process in a simple ML model is divided into two steps: training and testing. During the training process, samples from the dataset are taken as inputs whose characteristics correspond to a label. This data helps to construct the function model that represents the relationship between attributes and corresponding tags. During the testing process, the model takes as input the other characteristics, i.e. those that were not used in training, and through the function predicts their labels with some degree of success[24].

Supervised learning is the most common technique in classifying problems since its goal is for the machine to learn to classify a set of data that has been created. This set consists of features and tags. The main purpose is to build an appraiser, capable of predicting the label of an object given by the set of characteristics. The algorithm initially follows the training process by taking the inputs along with the corresponding correct tags and creates the classification model. Then follows the evaluation of the model, with the introduction of new data with known labels. These labels are compared with the corresponding model outputs and the prediction error and model accuracy are calculated. The above procedure is the second step, that of testing. Then, the model is modified accordingly. The model that is created functions correctly as long as there are available inputs, however, it cannot produce an output if any of the inputs are missing [23]. During this process in supervised learning, there is the risk of overfitting. This means that the model has not learned to generalize well, that is, it achieves excellent results during the training process, but performs much lower with the test data. It struggles, therefore, to handle samples that differ from those it sees during training. Thus, during the process of finding the optimal model, overfitting must be taken into account [26].

2.1.2. Neural Networks

Artificial Neural Networks (ANNs) are a computational system that mimic the structure and function of biological neural networks, such as the brain. More specifically, they simulate the processing of input signals and the production of appropriate output signals. In the context of the brain, input signals generated by some external stimulus are received by the dendrites and processed. The output is then received by other neurons or an organ, such as muscles. Similarly, in ANNs, neuron nodes that receive a vector of variables as input perform computations and produce an output, which is transmitted to subsequent neurons or constitutes the final output of the system model [25].

ANNs are a powerful tool in computer science that can be used to model complex problems for which the exact functions describing them are unknown. This foundational concept extends to the development of Large Language Models (LLMs). LLMs, such as GPT-3, are built on deep neural network architectures that leverage vast amounts of text data to learn and understand language patterns, grammar, and context. These models use multiple layers of neurons to process and generate human-like text, making them capable of tasks such as language translation, text generation, and sentiment analysis. By applying the principles of ANNs, LLMs are able to achieve a high degree of proficiency in natural language processing, showcasing the advanced capabilities of artificial neural networks in modeling and understanding complex data [27].

2.1.3. Evaluation metrics

After applying machine learning algorithms, some tools are needed to be able to understand how well they work. These tools are called performance evaluation metrics. There are a large number of metrics identified in studies, and each examines certain aspects of the algorithm's performance. Thus, for each machine learning problem, it is necessary to define the appropriate set of metrics [25].

In order to understand the following definitions, some terms should be given and explained. Specifically, if there is a problem with two classes, positive and negative, then true positives (TP) are defined as the samples that have been predicted in the positive class and actually belong to it. Similarly, true negatives (TN) are defined as samples that have been predicted in the negative class and actually belong to it. There is also the term false positive (FP), which reflects the number of samples predicted in the positive class, while belonging to the negative. Finally, false negative (FN) is similarly defined as the number of samples predicted in the negative class while belonging to the positive class.

In this paper, we use some of the most common metrics to get valuable insights into the performance of algorithms and compare them. The metrics we use are the following:

- Precision: Shows the number of selected data that is actually relevant. That is, of the observations that an algorithm has predicted belong to a class, how many of them actually belong to that class. According to the formula, precision equals the number of true positive (TP) observations, divided by the sum of true positives and false positives (FP).

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

- Recall: Represents the percentage of correct predictions by the algorithm for a class in terms of the set of observations belonging to it. According to the formula, recall is equal to the number of true positives (TP) divided by the sum of true positives and false negatives (FN).

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

- F1-score: This metric takes into account both precision and recall to calculate the performance of an algorithm. Mathematically, it equals the harmonic through these two metrics and is given by:

$$\mathrm{F1-score} = 2 \times \frac{\mathrm{Precision} + \mathrm{Recall}}{\mathrm{Precision} \times \mathrm{Recall}}$$

- Accuracy: It is the most used and perhaps the first choice for evaluating the performance of a classification algorithm. It can be defined as the percentage of correctly classified data in the total of observations. Of course, in some cases, it is not the most appropriate metric. This happens when the classes of the model are not equilibrium, i.e., we do not have the same number of observations for all classes. The formula is as follows*:*

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

- Confusion matrix: A confusion matrix summarizes the performance of a classifier with respect to certain test data. It is a two-dimensional array, which in one dimension consists of the actual class of an object and in the other consists of the class that the classifier assigns to that object. Figure 3 presents an example of a confusion table for a classification of three classes, A, B, C. The first row of the table shows that 13 objects

belong to class A and 10 are correctly classified to this class, two incorrect objects belong to B and one incorrect one normally belongs to C.

		Assigned Class		
		A	B	C
Actual Class	A	10	2	1
	B	0	6	1
	C	0	3	8

Figure 3: Example of 3 a confusion matrix of 3 classes [25]

In the case of 2-class problems, if one class is considered positive and the other negative, the table contains four positions within which are true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) [25]. The corresponding example is shown in Figure 4.

		Assigned Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Figure 4: Example of a confusion table with classes positive and negative [25]

2.2. Natural Language Processing (NLP)

Natural Language Processing (NLP) is likely to be the prime tool in linguistics and Artificial Intelligence which assists machines to figure out the human language and also to find something valuable in it. The implementation of ML and especially of NLP mental health monitoring, especially in social network analysis, is a very clever and innovative way how mental health providers to illuminate and uncover mental disorders like depression.

2.2.1. NLP and its Relevance to Mental Health Monitoring

Natural Language Processing (NLP) has emerged as a pivotal technology in enhancing mental health monitoring by enabling the analysis of language used across diverse digital platforms, such as social media and electronic health records.

Zhang [19] underscores the effectiveness of NLP in managing large volumes of unstructured text data, which is typically challenging for human supervision alone. This capability of NLP not only facilitates the early detection of mental health issues but also significantly improves the potential outcomes of subsequent treatments.

Nijhawan [28] discusses the transformative impact of NLP applications within the realm of psychology. Through techniques like sentiment analysis, emotion perception, and topic extraction, NLP offers a non-invasive and extensive method for tracking and evaluating individuals' mental states over time. This allows healthcare professionals to provide customized treatments that evolve with the patient's psychological condition.

The diagnostic of depressive themes through social media and other texts is also complicated because of the intricate NLP methods that are trying to work out the sensitive and ambiguous linguistic cues that are commonly used among those with mental conditions. Coppersmith [21] contributes to this area by integrating ML with NLP, enhancing the precision and efficiency of depression diagnosis. By training models such as Support Vector Machines

(SVM) and deep neural networks like Convolutional Neural Networks (CNN) on annotated text datasets, these techniques can effectively recognize depression indicators from linguistic patterns and word usage.

Further, advanced NLP models like Long Short-Term Memory (LSTM) networks, highlighted by Neelavathi [29], effectively capture the sequential order of words within texts, allowing for a more nuanced representation of emotions and ideas. This approach is crucial for differentiating between transient depressive episodes and persistent depressive symptoms, which require careful medical evaluation. Neelavathi also explores the use of sophisticated word embedding methods such as Word2Vec and GloVe, which enhance the sensitivity of NLP models to subtle language variations indicative of emotional states like loneliness and despair.

Overall, the integration of NLP into mental health monitoring represents a significant advancement in the field, providing essential tools for the early detection and ongoing management of mental health conditions. This chapter will delve deeper into how NLP technologies are being leveraged to improve mental health interventions and the quality of care provided to individuals suffering from mental health issues.

2.2.2. NLP Techniques and Metrics Used for Detecting Depressive Content

In the systematic review "Machine Learning for Depression Detection on Web and Social Media" conducted by Lin Gan, Yingqi Guo, and Tao Yang [20], the researchers meticulously analyzed the application of various machine learning algorithms and their evaluation metrics in the field of depression detection across web and social media platforms. This analysis aimed to discern which computational techniques are most prevalent and effective in identifying signs of depression.

The study highlighted a range of algorithms frequently used for this purpose. The Support Vector Machine (SVM) and Decision Tree (DT) algorithms were noted for their robustness in handling diverse data sets, reflecting traditional approaches to machine learning. Advanced neural network models such as the Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) were also prevalent, indicating a shift toward models capable of processing complex patterns in textual data. Additionally, ensemble methods like Random Forests (RF) and Gradient Boosting Machines (LightGBM and XGBoost) were recognized for their high accuracy and ability to mitigate overfitting by combining multiple decision trees. The review also acknowledged the relevance of Naive Bayes (NB) and K-Nearest Neighbors (KNN) in this domain. In general, there is a growing trend in NLP-driven research for detecting mental illnesses, highlighting the significant research value and potential for automated mental illness detection from text (Figure 5). Additionally, deep learning-based methods have gained popularity over the past few years [19].

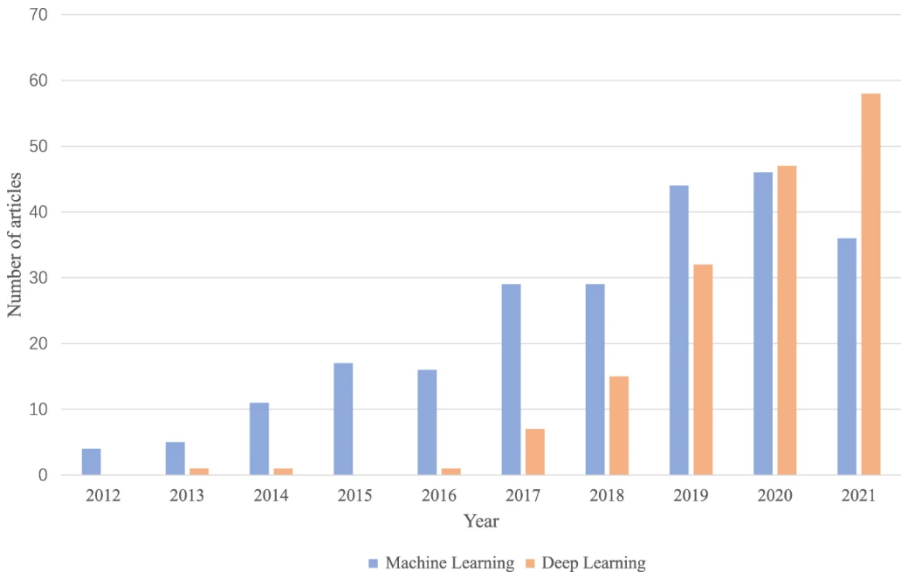

Figure 5: NLP trends applied to mental illness detection reseach using machine learning and deep learning [19]

In terms of natural language processing, the utilization of Bidirectional Encoder Representations from Transformers (BERT) and its derivatives underscored the importance of contextually rich models in interpreting the nuances of language used in social media posts. Other sophisticated techniques mentioned included Multilayer Perceptron (MLP), Maximum Entropy (ME), and novel approaches like the attention-based MFM-Att, highlighting the diversity of strategies employed to enhance depression detection.

Regarding the metrics used to evaluate these models, Accuracy and F1-score were the most commonly reported, emphasizing their importance in assessing the overall effectiveness and balance between precision and recall of the detection models. Other crucial metrics such as Precision and Recall were frequently used to ensure that models not only identify depressive content accurately but also minimize false negatives and false positives, which are critical in clinical settings. Advanced metrics like the Area Under the ROC Curve (AUC-ROC) and Matthews Correlation Coefficient (MCC) were also employed to provide deeper insights into the models' performance, especially in distinguishing between classes and handling imbalanced data.

This thorough analysis highlights the dynamic evolution of machine learning applications in mental health, illustrating the growing complexity and sophistication of methods and metrics used to tackle the challenge of depression detection in digital communications. The study provides a foundation for future research, emphasizing the need for continuous improvement in model accuracy, explainability, and ethical considerations in deploying these technologies[20].

2.3. Large Language Models (LLMs)

Large Language Models (LLMs) have become a groundbreaking innovation in NLP, providing new capabilities for understanding and generating human language. These models, such as GPTs and BERT, hold significant promise for improving depression detection systems by accurately identifying linguistic patterns indicative of mental health issues.

LLMs are a recent advancement in the NLP field achieved through transformers, increased computational resources, and vast amounts of training data. LLMs are capable of processing and generating human-quality text, and can be applied to a wide range of NLP tasks including translation, summarization, information retrieval, and conversational interactions. LLMs represent the culmination of advancements in NLP, transitioning from statistical language models to neural language models, pre-trained language models (PLMs), and finally to LLMs. PLMs are trained on large text corpora in a self-supervised manner to learn a general

representation for various NLP tasks. LLMs significantly outperform PLMs by leveraging a much larger number of model parameters and training data [30].

The emergence of LLMs has led to a surge in research, with numerous models being proposed in recent years. Early LLMs, such as T5 and mT5 (Figure 1), relied on transfer learning, whereas later models like GPT-3 demonstrated zero-shot transfer learning capabilities, eliminating the need for fine-tuning on specific tasks. However, pre-trained LLMs can struggle with user intent and perform better in few-shot settings compared to zero-shot settings. Fine-tuning LLMs with task instruction data and aligning them with human preferences can significantly improve their performance in zero-shot settings and reduce misaligned behavior.

Beyond enhanced generalization and domain adaptation, LLMs exhibit emergent abilities such as reasoning, planning, decision-making, and in-context learning, even though these skills are not explicitly trained. These capabilities have led to the adoption of LLMs in diverse applications including multi-modal domains, robotics, tool manipulation, question-answering, and autonomous agents.

Despite their impressive capabilities, LLMs come with limitations such as slow training and inference times, significant hardware requirements, and high operational costs. These limitations hinder widespread adoption and motivate research into more efficient LLM architectures, training strategies, and parameter reduction techniques [30].

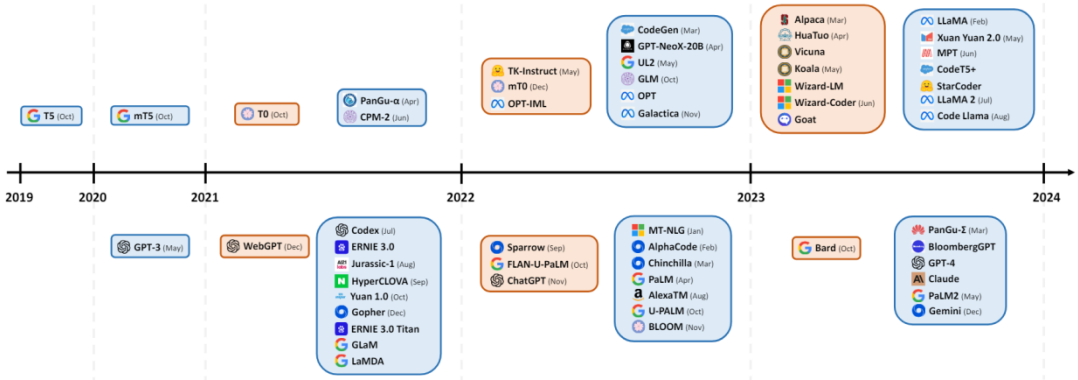

Figure 6: Chronological display of LLM releases: blue cards represent 'pre-trained' models, while orange cards correspond to 'instruction-tuned' models. Models [30]

2.3.1. LLMs and Mental Health Prediction

NLP methods have become a valuable tool for analyzing mental health using social media content. This chapter explores the evolution of NLP techniques, with a particular focus on the recent advancements brought about by LLMs.

Early NLP methods primarily relied on traditional word embedding techniques like Word2Vec and GloVe. While these methods offered a way to represent words numerically, they often lacked the ability to capture the full semantic context, particularly for ambiguous language found in social media. This limitation hindered the accuracy of mental health classification tasks.

The introduction of LLMs, such as BERT, marked a significant leap forward in NLP's ability to process and understand natural language. LLMs leverage powerful neural network architectures to achieve superior performance in tasks like sentiment analysis and text classification. However, research suggests that LLMs may still require fine-tuning for optimal performance in specific domains like mental health analysis.

A key concept in NLP for mental health analysis is the use of embeddings. Embeddings represent words as vectors in a multi-dimensional space, where the position of a word within the space reflects its semantic and syntactic similarities to other words. This allows NLP models to capture nuanced relationships between words, which is crucial for identifying mental health indicators within social media text.

Studies have explored various embedding techniques, including those generated by LLMs like BERT, alongside multi-level embeddings that combine representations from different linguistic units. These advancements have shown promise in improving the accuracy of mental health prediction tasks. For instance, research has demonstrated that combining embeddings with classifiers like Long Short-Term Memory (LSTM) networks can achieve high accuracy in classifying social media content related to mental health.

The potential of LLMs and advancements in NLP offer exciting possibilities for large-scale mental health analysis. By leveraging the vast amount of data available on social media platforms, researchers can gain valuable insights into population-level mental health trends. This information can be used to develop targeted interventions and improve mental health support systems.

However, challenges remain in utilizing NLP for mental health analysis. Robust validation methods are needed to ensure the accuracy and generalizability of the findings. Additionally, ethical considerations regarding user privacy and responsible data usage must be addressed to ensure the responsible implementation of these technologies [31].

2.3.2. Chatbots for Mental Health Support

Mental health chatbots represent a burgeoning intersection of artificial intelligence (AI) and mental health care, leveraging the capabilities of LLMs to provide accessible, scalable, and personalized support. These chatbots, designed to simulate human conversation, utilize advanced AI techniques, including NLP and ML, to interact with users, offering real-time assistance and mental health interventions[32].

LLMs such as GPT-3 and BERT have significantly enhanced the functionality of mental health chatbots. These models are trained on vast datasets and are capable of understanding and generating human-like text, which enables chatbots to engage in meaningful conversations with users. By leveraging LLMs, mental health chatbots can understand context, detect emotional cues, and provide responses that are both relevant and empathetic [33].

The integration of LLMs into mental health chatbots has facilitated the development of sophisticated tools that can analyze user inputs, recognize signs of mental distress, and offer appropriate guidance or resources. These chatbots can perform tasks ranging from conducting preliminary mental health assessments to providing cognitive-behavioral therapy (CBT) techniques and emotional support [34].

Woebot [35] is a prominent mental health chatbot that uses Cognitive Behavioral Therapy (CBT) techniques to help users manage symptoms of depression and anxiety, including postpartum depression. Woebot initiates daily therapy sessions, asking users about their feelings and providing CBT-based activities or challenges. The chatbot aims to mimic human conversation to offer emotional support and practical advice. Research funded by Woebot's creators indicates promising results in reducing depressive symptoms over short periods, though independent validation is required for more robust conclusions.

Wysa [36] utilizes CBT and other therapeutic techniques to provide a safe, non-judgmental space for users to discuss their concerns. It offers self-help tools for reframing issues, educational content on various mental health topics, and access to professional therapists through a premium subscription. Wysa's emphasis on privacy and security ensures that user conversations remain confidential, addressing one of the significant concerns associated with digital mental health interventions.

Youper [37] integrates CBT and Positive Psychology techniques to help users manage anxiety and depression. The app features a conversational AI that provides just-in-time interventions and personalized recommendations based on user interactions. Founded by doctors and therapists, Youper emphasizes clinical effectiveness and continuous improvement, making it a reliable tool for those seeking evidence-based mental health support.

Coach Marlee [38], developed by Fingerprint for Success (F4S), is an AI-powered coach designed to enhance users' mental performance and well-being. Based on over 20 years of scientific research, Marlee offers personalized coaching programs that help users achieve their goals and improve their mental health. The chatbot provides friendly check-ins, motivational content, and a platform for self-reflection and growth. While Marlee focuses on behavioral change rather than therapeutic intervention, it can be a valuable tool for developing emotional resilience and self-esteem.

The global shortage of mental health professionals has driven the adoption of mental health chatbots. These digital tools offer affordable, accessible support, particularly for individuals who face barriers to traditional therapy such as cost, availability, and stigma. Research indicates that chatbots can effectively provide emotional support, improve access to care, and reduce the burden on mental health services .

Mental health chatbots are particularly beneficial for those who prefer anonymity and immediate access to support. Studies have shown that some users are more willing to disclose personal information to chatbots than to human therapists, highlighting the potential for these tools to reach individuals who might otherwise avoid seeking help.

Several studies have examined the effectiveness of mental health chatbots. For example, a randomized controlled trial of Woebot demonstrated significant reductions in depressive symptoms within two weeks of use. Similarly, evaluations of Wysa and Youper have shown positive outcomes in managing anxiety and depression, with users reporting high satisfaction and perceived improvement in their mental health.

However, the research also highlights some limitations. Chatbots are not suitable for crisis intervention or severe mental health conditions, as they lack the ability to provide nuanced, real-time human empathy and support. Privacy concerns and the need for robust data handling practices are also critical issues that developers must address to ensure user trust and safety [34].

Research on mental health chatbots typically involves collecting data from users through surveys, user interactions, and clinical trials. For instance, studies on Woebot and Wysa have used datasets comprising user interactions, feedback, and self-reported mental health outcomes to evaluate effectiveness. These datasets help researchers understand user behavior, refine chatbot algorithms, and improve the overall user experience [34].

In conclusion, mental health chatbots represent a significant advancement in digital mental health support, offering accessible and cost-effective alternatives to traditional therapy. While they are not a replacement for professional mental health care, they provide valuable support for individuals with mild to moderate symptoms of mental health issues. Ongoing research and development are crucial to enhance their effectiveness, address ethical concerns, and ensure they complement existing mental health services effectively.

2.4. Datasets, Approaches, Ethical Considerations and Regulatory in Mental Health Monitoring

2.4.1. Creation and Utilization of Datasets

The performance of an NLP system in recognizing depression in social media posts heavily relies on the quality and scope of the datasets used. Developing robust databases is crucial when training machine learning models capable of identifying language patterns associated with depression. The process involves meticulous collection, annotation, and refinement of data, which several researchers have extensively explored.

The creation of these datasets typically involves several key steps. Data collection is the initial step, where large volumes of text data are gathered from social media platforms. This often involves scraping posts from forums, blogs, and other online communities where users discuss their mental health experiences. Annotation follows, where the collected data is

labeled to indicate the presence or absence of depressive symptoms. This step is crucial for training supervised learning models, and it is often performed by expert annotators or crowd-sourced workers. Preprocessing involves cleaning the text data to remove noise, such as irrelevant content, advertisements, or spam. It also includes tokenization, lemmatization, and other text normalization techniques to prepare the data for analysis. Feature extraction is another important step, using NLP techniques to extract relevant features from the text, such as sentiment scores, emotional tone, and linguistic patterns. Advanced models like BERT and GPT-3 can capture contextual embeddings that provide deeper insights into the text. Finally, model training and validation involve training machine learning models on the annotated datasets and validating their performance using metrics such as accuracy, precision, recall, and F1-score.

The use of these datasets extends beyond the identification of individuals with mental health issues. They are also employed for longitudinal analysis, allowing researchers to track changes in personal behavior and communication styles over time. This analysis can reveal trends in psychological status, providing valuable insights for targeted and preventive mental health interventions.

Several popular datasets have been utilized in depression prediction research. One notable dataset is the "Reddit Self-reported Depression Diagnosis" dataset, which consists of posts from individuals who have self-identified as experiencing depression. This dataset is particularly valuable due to its extensive annotations and the rich linguistic patterns it contains, providing a robust foundation for training NLP models [39]. Another significant dataset is the "CLPsych 2015 Shared Task Dataset," which includes a collection of posts from an online mental health forum. This dataset has been used extensively in research due to its well-documented annotations and the depth of information it offers on mental health discourse [40]. Additional datasets include the "eRisk Dataset", which is part of the CLEF eRisk challenge series. This dataset includes a variety of social media posts that have been annotated for signs of mental health issues, including depression. It has been widely used for benchmarking NLP models in depression detection tasks [41]. Finally, an additional dataset is the "Twitter Depression Dataset", which comprises tweets labeled for depressive content. This dataset is valuable for studying the expression of depressive symptoms in short, informal text formats typical of Twitter [42].

Generally, datasets from social media text for depression prediction use at least two key features: the text from users and the corresponding label indicating whether the content is associated with depression or not. These features are essential for training and evaluating models to ensure they can accurately identify depressive language patterns.

2.4.2. Most recent approaches in mental Health prediction

Recent advancements [43] in LLMs have introduced significant improvements in the ability to predict mental health issues through online text data [43]. LLMs, including models such as Alpaca, Alpaca-LoRA, FLAN-T5, GPT-3.5, and GPT-4, have shown considerable promise in accurately identifying linguistic patterns indicative of mental health problems like stress, depression, and suicidal ideation. This chapter delves into the methodologies employed in these LLMs, focusing on their application in mental health prediction tasks.

Large Language Models (LLMs) such as GPT-3 and BERT represent a transformative advancement in Natural Language Processing (NLP). These sophisticated models are trained on vast datasets and leverage deep learning techniques to understand and generate human language, capturing intricate details of context, semantics, and syntax. The exceptional performance of LLMs in various NLP tasks—including text generation, translation, summarization, and sentiment analysis—has opened new frontiers in numerous applications, particularly in mental health monitoring.

In mental health monitoring, LLMs offer powerful capabilities for analyzing large volumes of unstructured text data from social media posts, electronic health records, and other digital communications. Their ability to discern subtle nuances in language makes them particularly effective for detecting early signs of depression. By meticulously analyzing the patterns in which individuals express their thoughts and emotions, LLMs can provide invaluable insights for early intervention and continuous mental health monitoring.

The evaluation of LLMs for mental health prediction tasks relies on various well-established datasets that provide a diverse range of mental health-related text data. One such dataset is Dreaddit, which contains posts from Reddit collected via the Reddit PRAW API between January 1, 2017, and November 19, 2018. It includes posts from ten subreddits related to abuse, social anxiety, PTSD, and financial stress. Human annotators rated whether the sentence segments showed stress, and these annotations were aggregated to generate final labels. This dataset was used for post-level binary stress prediction.

Another dataset, DepSeverity, utilizes the same posts as Dreaddit but focuses on depression. This dataset categorizes posts into four levels of depression: minimal, mild, moderate, and severe, based on DSM-5 criteria. It was employed for binary and four-level depression prediction tasks. The SDCNL dataset comprises posts from Reddit's r/SuicideWatch and r/Depression, collected from 1,723 users. Posts were manually annotated for suicidal thoughts, and it was used for post-level binary suicide ideation prediction.

Additionally, the CSSRS-Suicide dataset, collected from 15 mental health-related subreddits from 2,181 users, includes manual annotations by psychiatrists following the Columbia Suicide Severity Rating Scale (C-SSRS) guidelines. This dataset was utilized for user-level binary and five-level suicide risk prediction. The Red-Sam dataset, also from Reddit, includes posts from subreddits related to mental health, depression, loneliness, stress, and anxiety. It was used for external evaluation of binary depression detection.

The Twt-60Users dataset, collected from Twitter, includes tweets from 60 users during 2015, labeled for depression by human annotators. It served as an external evaluation dataset for binary depression detection. Lastly, the SAD dataset contains SMS-like text messages written by 3578 humans under stressor-triggered instructions, labeled for various types of daily stressors. It was used for external evaluation of binary stress detection.

The study involves multiple LLMs with different configurations and pre-training targets. Alpaca (7B) is an open-source model fine-tuned from LLaMA 7B on instruction-following demonstrations. It represents a compact yet powerful model designed for various NLP tasks. Alpaca-LoRA (7B) is similar to Alpaca but fine-tuned using low-rank adaptation to reduce finetuning cost. This technique freezes the model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture. FLAN-T5 (11B) is an open-source model focused on task-solving, finetuned with a variety of task-based datasets. FLAN-T5 prioritizes task-specific performance over general language generation.

LLaMA2 (70B) is a recent model released by Meta, picked for its large size and advanced capabilities. GPT-3.5 (175B) is a closed-source model available through OpenAI's API, renowned for its extensive training on diverse datasets and exceptional performance in language understanding tasks. GPT-4 (1700B) is the largest model by OpenAI, offering unparalleled language comprehension and generation capabilities, albeit with significant computational requirements.

The experimental setups employed in the studies include zero-shot prompting, few-shot prompting, and instruction finetuning. Zero-shot prompting involves evaluating the models without domain-specific training by using carefully designed prompts tailored for mental health tasks. The general zero-shot prompt template consists of the online text data (TextData), a specification for a mental health prediction target (PromptPart1-S), a question for the LLMs to answer (PromptPart2-Q), and an output constraint (OutputConstraint).

Few-shot prompting enhances the models' performance by providing a few examples within the prompts to facilitate domain-specific learning. The few-shot prompt setup involves adding additional prompt-label pairs to the original prompt template. Instruction finetuning involves finetuning the models on multiple mental health datasets to optimize their performance for specific mental health tasks. This process involves updating the model parameters based on a small amount of domain-specific data.

The results indicate that while zero-shot and few-shot prompting show promising performance, instruction finetuning significantly enhances the models' capabilities across multiple tasks. Mental-Alpaca and Mental-FLAN-T5, the finetuned versions, outperform the best zero-shot and few-shot prompt designs of GPT-3.5 and GPT-4 in balanced accuracy. They also perform on par with the task-specific state-of-the-art model, Mental-RoBERTa [24].

Dataset	*Description*	*Social Media Used*	*Scope*
Dreaddit	Contains posts from Reddit collected via the Reddit PRAW API between January 1, 2017, and November 19, 2018. Includes posts from ten subreddits related to abuse, social anxiety, PTSD, and financial stress. Human annotators rated stress levels, aggregated to generate final labels.	Reddit	Post-level binary stress prediction
DepSeverity	Utilizes the same posts as Dreaddit but focuses on depression. Categorizes posts into four levels of depression: minimal, mild, moderate, and severe, based on DSM-5 criteria.	Reddit	Binary and four-level depression prediction
SDCNL	Comprises posts from Reddit's r/SuicideWatch and r/Depression, collected from 1,723 users. Manually annotated for suicidal thoughts.	Reddit	Post-level binary suicide ideation prediction
CSSRS-Suicide	Collected from 15 mental health-related subreddits from 2,181 users. Includes manual annotations by psychiatrists following the Columbia Suicide Severity Rating Scale (C-SSRS) guidelines.	Reddit	User-level binary and five-level suicide risk prediction
Red-Sam	Includes posts from subreddits related to mental health, depression, loneliness, stress, and anxiety.	Reddit	External evaluation of binary depression detection

Twt-60Users	Collected from Twitter, includes tweets from 60 users during 2015, labeled for depression by human annotators.	Twitter	External evaluation of binary depression detection
SAD	Contains SMS-like text messages written by 3,578 humans under stressor-triggered instructions, labeled for various types of daily stressors.	SMS	External evaluation of binary stress detection
Reddit Self-reported Depression Diagnosis	Posts from individuals who have self-identified as experiencing depression, extensively annotated for training NLP models.	Reddit	Identification and analysis of depressive language patterns
CLPsych 2015 Shared Task Dataset	Collection of posts from an online mental health forum, well-documented annotations providing depth of information on mental health discourse.	Online Forum	Analysis of mental health discourse
eRisk Dataset	Part of the CLEF eRisk challenge series, includes various social media posts annotated for signs of mental health issues, including depression.	Mixed (Social Media)	Benchmarking NLP models in depression detection tasks
Twitter Depression Dataset	Tweets labeled for depressive content, valuable for studying the expression of depressive symptoms in short, informal text formats typical of Twitter.	Twitter	Studying depressive symptoms in short, informal text formats

Figure 7: Common datasets used in Mental Health prediction

Model	***Description***
Alpaca (7B)	An open-source model fine-tuned from LLaMA 7B on instruction-following demonstrations. Compact yet powerful, designed for various NLP tasks.
Alpaca-LoRA (7B)	Similar to Alpaca but fine-tuned using low-rank adaptation to reduce finetuning costs. This technique freezes the model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture.
FLAN-T5 (11B)	An open-source model focused on task-solving, finetuned with a variety of task-

	based datasets. Prioritizes task-specific performance over general language generation.
LLaMA2 (70B)	A recent model released by Meta, selected for its large size and advanced capabilities.
GPT-3.5 (175B)	A closed-source model available through OpenAIs API, renowned for extensive training on diverse datasets and exceptional performance in language understanding tasks.
GPT-4 (1700B)	The largest model by OpenAI, offering unparalleled language comprehension and generation capabilities with significant computational requirements.
BERT	Bidirectional Encoder Representations from Transformers (BERT) is an open-source model pre-trained on a large corpus of text. It uses bidirectional training to understand the context of words in a sentence, making it particularly effective for a variety of NLP tasks, including text classification, sentiment analysis, and named entity recognition. It is widely used in research and practical applications.

Figure 8: LLMs used for Mental Health prediction

2.4.3. Ethical Considerations and Challenges in Data Collection

Privacy and consent issues often become the first ethical concerns discussed in mental health monitoring when social media platforms serve as the data sources. In Kim's [17] argument, regulation should be tightened for data anonymity and encryption procedures. It is emphasized that it is not sufficient to merely anonymize datasets; all personal information must be thoroughly removed before data is saved for any further use. This approach ensures that data is protected against unauthorized access. Kim further stresses the importance of clear data collection procedures where users are fully informed about how their data will be used, thereby ensuring that their privacy and autonomy are respected.

Ethical considerations in using social media data for health and healthcare research are critical. The dramatic growth of social media usage necessitates clear guidelines and frameworks to ensure privacy, confidentiality, and ethical data handling [44]. Research by Hunter [45] outlines key ethical concerns for public health researchers using social media, including privacy, anonymity, data security, and management. The ethical use of social media data extends beyond the clinical context, and guidelines must be adapted to research-specific issues to ensure the ethical conduct of research.

Malhotra and Jindal [46] broaden this discourse by examining the ethical implications of using social media data for mental health purposes. They highlight the potential for privacy violations if online data is utilized without proper consent. Their strategy involves formulating rules to balance the ethical considerations of social media data operations with privacy protection measures, which include informed consent, data minimization, and governance frameworks that control data usage

Moreover, ethical problems become more intricate depending on the specifics of data collection. Social media users do not represent a homogeneous group; they vary widely in age, behavior, and online engagement. This diversity suggests that datasets may not be representative of the general public, leading to the development of biased models. To mitigate

these biases, it is crucial to involve diverse viewpoints during data collection, ensuring the developed models are robust and unbiased. Continuous and iterative feedback loops during data collection and model training are necessary to enhance the model's fairness and accuracy.

Another significant ethical challenge is the potential psychological impact on individuals whose data is being used. Researchers must consider the implications of their findings and the potential for stigmatization or unintended harm. Ethical guidelines must be established to address these concerns, ensuring that the research benefits outweigh the risks to individuals. The use of social media data in mental health research can lead to issues such as incorrect, opaque algorithmic predictions, involvement of bad or unaccountable actors, and potential biases from intentional or inadvertent misuse of insights[10].

2.4.4. Regulatory and Compliance Issues

The use of social media data for health observation operates within a complex regulatory landscape that varies widely across different jurisdictions. The General Data Protection Regulation (GDPR) of the European Union (EU) imposes strict requirements on the handling of personal data, including the necessity of obtaining explicit consent from users and ensuring their data is anonymized and secure. Researchers and practitioners must comply with these regulations, which govern the collection, processing, and storage of personal data [45]. The GDPR ensures that personal data is handled with the highest standards of privacy and security, safeguarding individuals' rights and freedoms in the digital age.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) provides a regulatory framework for the protection of health information. While HIPAA primarily addresses medical records, its principles can be applied to the handling of social media data used for health research, emphasizing the importance of privacy and security

Additionally, the ethical framework for research involving human subjects, such as the Belmont Report, outlines principles of respect for persons, beneficence, and justice. These principles should guide the ethical conduct of research involving social media data, ensuring that participants are treated with dignity, their welfare is prioritized, and the benefits and burdens of research are distributed fairly (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). The Belmont Report's principles are foundational to ethical research practices, ensuring that all human subjects are protected and respected throughout the research process.

Researchers must navigate these regulatory frameworks with meticulous attention to detail and ethical practice. This includes obtaining institutional review board (IRB) approval for studies involving human subjects, conducting thorough risk assessments, and implementing robust data protection measures[47]. The IRB approval process is crucial for evaluating the ethical implications of the research and ensuring that all procedures comply with established ethical standards.

Moreover, the ethical use of social media data extends beyond compliance with regulatory standards. Ethical considerations must address the potential risks and benefits associated with using social media data for health research. This involves implementing best practices in data anonymization, informed consent, and bias mitigation. Ethical research protocols, such as those suggested by Hunter [45], emphasize the need for transparency, accountability, and respect for privacy when using social media data.

In summary, the process of developing and employing datasets for depression diagnosis through NLP requires a comprehensive approach that addresses both technical and ethical aspects. By ensuring compliance with regulatory standards and adopting best practices in data anonymization, informed consent, and bias mitigation, researchers can responsibly harness the potential of social media data to advance mental health monitoring and intervention.

2.5. LLMs and Depression Prediction

This chapter delves into the specific methodologies, results, and findings of key studies that have applied LLMs to predict depression based on social media posts.

One significant study by Xin [48] aimed to enhance understanding of depression treatment outcomes by incorporating broader measures beyond just depressive symptoms, focusing on the quality of life and interpersonal relationships. Researchers used data from the IMPACT-My Experience study, involving interviews with adolescents, parents, and therapists. These interviews were annotated using a comprehensive framework identifying specific outcomes. LLMs, including BERT, MentalBERT, MentalLongformer, and Llama 2-7B, were used to classify these outcomes from the text. The study found that Llama 2-7B outperformed other models, and domain-specific models like MentalBERT showed improved performance. Different text segmentations impacted the models' effectiveness, with monologue and turn segmentations yielding better results.

The study demonstrated that these models could effectively automate the identification of outcomes relevant to depression treatment, achieving robust performance metrics. Additionally, the research highlighted the potential of LLMs to predict depression-related outcomes by capturing nuanced linguistic patterns indicative of depressive symptoms. This capability underscores the potential of LLMs to serve as early warning systems, enabling timely interventions that could mitigate the severity of depression in adolescents. The findings emphasize the importance of leveraging advanced NLP techniques to enhance the precision and depth of mental health research, ultimately contributing to more effective mental health strategies and interventions.

In another pivotal research effort, De Duro [49] introduced the CounseLLMe dataset, a novel collection of simulated mental health dialogues to compare the performance of different LLMs, including Haiku, LLaMAntino, and ChatGPT. The focus of the study was on evaluating the models' ability to simulate realistic mental health counselling sessions, capturing the nuances of patient-therapist interactions. The researchers found that LLMs like ChatGPT could realistically reproduce several aspects of human conversations, including emotional exchanges and syntactic patterns, though they struggled to replicate certain emotional responses such as anger and frustration. This study is significant as it demonstrates the potential of LLMs to function as tools in mental health support systems by providing realistic simulations of therapeutic dialogues. The findings highlight the applicability of LLMs in scenarios requiring human-like understanding and empathy, showcasing their potential to aid in mental health interventions and training

In the study [50] Elyoseph and Levkovich, the researchers explored the effectiveness of LLMs, including ChatGPT-3.5 and ChatGPT-4, in evaluating the prognosis of depression when compared with assessments from mental health professionals and the general public. The research highlighted how LLMs could simulate professional prognostic assessments, often aligning closely with or differing from human evaluators in their predictions of long-term outcomes for patients with depression. The study discovered variations in the optimism or pessimism of the model's prognoses, notably that ChatGPT-3.5 often provided more pessimistic forecasts compared to other LLMs and human assessors. This comparative analysis underscores the potential and limitations of using advanced AI in mental health settings to support clinical decision-making processes

In the research by Pérez [51] , details the creation of the DepreSym dataset, which consists of 21,580 sentences specifically annotated for relevance to the 21 depressive symptoms outlined in the Beck Depression Inventory-II (BDI-II). This dataset was compiled through a meticulous assessment methodology involving diverse ranking methods to identify sentences from online publications, followed by a detailed review process executed by three expert assessors, including a clinical psychologist. The primary aim was to develop a robust

dataset that accurately reflects the various manifestations of depressive symptoms in text, providing a foundation for further research in depression detection. Additionally, the study investigated the potential of leveraging LLMs such as ChatGPT and GPT-4 to assist in this complex annotation task, aiming to combine the efficiency of machine learning with the nuanced understanding of human experts. The research findings indicated that while LLMs like ChatGPT and GPT-4 can significantly streamline the annotation process by pre-assessing the relevance of texts to specific depressive symptoms, they still require close supervision and verification by human annotators to ensure accuracy and clinical relevance. The expert assessors evaluated the LLMs' performance, identifying both strengths and limitations in their ability to mimic human-like assessment capabilities. Despite some challenges, the LLMs demonstrated a promising level of competence, suggesting their utility as supplementary tools in the annotation process. This integration of LLMs aims to enhance both the scalability and quality of datasets like DepreSym, which are crucial for advancing the development of automated systems capable of detecting depression from textual data on social media and other platforms.

The study "Mental-llm: Leveraging Large Language Models for Mental Health Prediction via Online Text Data" Xu [43] critically examines the application of various Large Language Models (LLMs) like Alpaca, Alpaca-LoRA, FLAN-T5, GPT-3.5, and GPT-4, focusing on their effectiveness in predicting mental health outcomes from online text data. This study, involving collaborations from top-tier institutions such as MIT, Stanford, and others, utilizes a methodological framework that includes zero-shot prompting, few-shot prompting, and instruction fine-tuning to evaluate these models.

The findings indicate that while initial results from zero-shot and few-shot prompting are promising, significant enhancements in performance are observed when models undergo instruction fine-tuning. Notably, the fine-tuned models, Mental-Alpaca and Mental-FLAN-T5, outperform their larger counterparts like GPT-3.5 and GPT-4 in balanced accuracy metrics. The research further explores the reasoning capabilities of these models in mental health assessments, particularly highlighting the potential of GPT-4 in this domain[43].

In addition to technical performance assessments, the study provides practical guidelines for improving LLM applications in mental health tasks and emphasizes the importance of ongoing model refinement. It also addresses the critical ethical concerns and biases inherent in deploying LLMs for sensitive applications such as mental health diagnostics. The research calls for the development of strategies to mitigate these risks, ensuring that the deployment of such models in real-world settings is both responsible and effective [43].

In the study titled "Depression Detection on Social Media with Large Language Models" by Lan et al. [40], the researchers introduced a sophisticated system called DORIS that combines Large Language Models (LLMs) with traditional classifiers to improve the detection of depression from social media posts. This system uniquely leverages professional medical knowledge by annotating high-risk texts according to medical diagnostic criteria and summarizing users' mood histories to enhance predictive accuracy. DORIS integrates these features with Gradient Boosted Trees, a method that contributes to the high accuracy and explainability of the results. This dual approach not only detects signs of depression but also provides explanations for its predictions, marking a significant advancement in the application of LLMs for mental health monitoring on social media platforms.

In the study by Farruque [52], a novel semi-supervised learning approach was employed to model depression symptoms from social media texts, focusing on capturing the nuanced expressions of various depression levels. Utilizing a combination of a supervised learning model and a zero-shot learning model, both fine-tuned on a clinician-annotated dataset, the research demonstrated effective detection of depression symptoms. This approach not only leveraged the large volume of data available on social media but also refined the model's accuracy through iterative retraining on newly harvested data, showcasing

the LLMs' capability to discern subtle linguistic cues associated with different depression states. The study underscores the potential of semi-supervised learning in enhancing the precision of mental health assessments, contributing significantly to the nuanced understanding and treatment of depression based on social media content.

Lastly, Ohse [53] investigated the use of Natural Language Processing (NLP) models for detecting depression from clinical interview transcripts, the authors evaluated four prominent NLP models: BERT, Llama2-13B, GPT-3.5, and GPT-4. The study involved 82 participants who underwent clinical interviews and completed self-report depression questionnaires. The models analyzed the transcripts to infer depression scores, which were then compared against questionnaire cut-off values to classify depression. The study found that GPT-4 had the highest accuracy for depression classification with an F1 score of 0.73. GPT-3.5, initially performing with low accuracy at 0.34, showed significant improvement to 0.82 accuracy after fine-tuning and maintained a 0.68 score with clustered data. Moreover, GPT-4's estimates of PHQ-8 symptom severity scores correlated strongly (r = 0.71) with actual symptom severity reported by participants. These results underscore the robust potential of AI models in the accurate detection of depression, signaling a promising future for AI application in clinical settings, although the study suggests that more research is needed before these technologies can be broadly deployed

Study	*Models Used*	*Dataset and Data Collection*	*Key Focus*	*Results*	*Future Work*
Xin [24]	BERT, MentalBERT, MentalLongformer, Llama 2-7B	IMPACT-My Experience study, interviews with adolescents, parents, and therapists; annotated text	Identify treatment outcomes beyond depressive symptoms	Llama 2-7B outperformed other models; monologue and turn segmentations yielded better results	Leveraging advanced NLP techniques to enhance precision and depth of mental health research
De Duro [25]	Haiku, LLaMAntino, ChatGPT	CounseLLMe dataset, simulated mental health dialogues	Evaluate LLMs in simulating mental health counseling sessions	ChatGPT realistically reproduced several aspects of human conversations, but struggled with certain emotional responses	Using LLMs in mental health support systems for realistic simulations of therapeutic dialogues

Elyoseph and Levkovich [26]	ChatGPT-3.5, ChatGPT-4	Evaluation of LLMs' prognostic assessments compared with mental health professionals and the general public	Compare LLMs' prognostic assessments to human assessments	ChatGPT-3.5 provided more pessimistic forecasts; LLMs closely aligned with professional assessments	Supporting clinical decision-making processes with advanced AI
Perez [27]	ChatGPT, GPT-4	DepreSym dataset, 21,580 sentences annotated for relevance to the Beck Depression Inventory-II	Develop a robust dataset for depressive symptom detection	LLMs like ChatGPT and GPT-4 streamlined the annotation process but required supervision	Enhancing the scalability and quality of depression detection datasets
Xu [19]	Alpaca, Alpaca-LoRA, FLAN-T5, GPT-3.5, GPT-4	Online text data	Evaluate effectiveness of various LLMs in mental health tasks	Fine-tuned models (Mental-Alpaca and Mental-FLAN-T5) outperformed larger models like GPT-3.5 and GPT-4	Improving LLM applications in mental health tasks and addressing ethical concerns
Lan et al. [40]	Gradient Boosted Trees, LLMs, traditional classifiers	Social media posts	Improve depression detection by combining LLMs and classifiers	High accuracy and explainability in depression detection by combining LLMs with traditional classifiers (DORIS system)	Advancing mental health monitoring on social media platforms
Farruque [28]	Supervised learning model, zero-shot learning model	Social media texts, clinician-annotated dataset	Model depression symptoms with semi-supervised learning	Effective detection of depression symptoms, refined model accuracy through iterative retraining on newly harvested data	Enhancing the precision of mental health assessments with semi-supervised learning
Ohse [29]	BERT, Llama2-13B, GPT-3.5, GPT-4	Clinical interview transcripts, self-report depression questionnaires	Detect depression from clinical interview transcripts	GPT-4 achieved the highest accuracy for depression classification with an F1 score of 0.73; significant improvement in GPT-3.5 after fine-tuning	Investigating further research before broad deployment of AI models in clinical settings

Figure 9: Summary of research with LLMs for depression prediction

2.6. NLP and Suicide Prediction

This chapter explores the application of Natural Language Processing (NLP) and Large Language Models (LLMs) in predicting suicidal ideation through social media interactions. It draws from a variety of studies, each employing specific NLP techniques to identify patterns indicative of suicide risk from the data generated on various social media platforms. The focus

is on detailing the methodologies employed, the datasets used, the platforms analyzed, and the findings of each study.

One significant contribution to this field is the study by Coppersmith [21] employed a variety of NLP techniques to analyze Twitter data and identify linguistic markers indicative of mental health issues, including suicidal thoughts. The methodologies used included Sentiment Analysis, which is commonly applied to analyze the emotional content of texts. This technique allows for the detection of mood states that may be indicative of mental health concerns. Additionally, advanced machine learning models, likely involving aspects of deep learning such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN), were utilized to handle the complexity and subtleties of language used in tweets expressing mental health concerns.

Aldhyani [22] took a deep dive into the Reddit platform, specifically analyzing posts from the SuicideWatch subreddit. The dataset included 232,074 posts, balanced between suicidal and non-suicidal content. The study combined Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory networks (BiLSTMs) for deep learning, alongside traditional machine learning models like XGBoost. CNNs were employed to extract spatial features from text data, while BiLSTMs captured temporal context, enhancing the model's ability to discern complex patterns indicative of suicidal ideation. The methodology proved highly effective, with the deep learning model showing remarkable accuracy in identifying posts that exhibited suicidal ideation.

Wang [54] expanded the scope of data sources by gathering data from multiple social media platforms to examine the emotional and psychological expressions of users. This study employed various LLMs to process the data, focusing on advanced sentiment analysis and the extraction of semantic and syntactic features, which are crucial for understanding the context and emotional undertones. The findings indicated that LLMs trained on diverse datasets could accurately predict suicidal tendencies, highlighting the crucial role of comprehensive data analysis in suicide prevention.

Fonseka [55] integrated data from clinical interviews and social media posts to create a richer picture of individuals' mental health. The study applied keyword extraction and thematic analysis to identify indicators of suicidal behavior from the combined datasets. The integration of clinical data provided an additional layer of validation for the NLP models used, enhancing the predictive accuracy concerning suicidal behaviors. The results suggested that merging various data sources could lead to significant improvements in the early detection of suicide risk.

Lastly, Tadesse [56] focused exclusively on Reddit data, specifically analyzing posts from mental health-related subreddits. They employed advanced deep learning techniques, particularly LSTM networks, which excel at capturing complex dependencies in sequence data. This approach allowed for a nuanced understanding of textual data, effectively identifying emotional and psychological states indicative of suicidal ideation.

In addition, Roy [57] developed the Suicide Artificial Intelligence Prediction Heuristic (SAIPH), an algorithm capable of predicting future risk to suicidal thoughts by analyzing Twitter data. This model trained on tweets related to psychological constructs such as burden, stress, loneliness, hopelessness, insomnia, depression, and anxiety, and achieved a high area under the curve (AUC) of 0.88 in predicting suicidal ideation events, demonstrating the algorithm's potential in identifying individual future SI risk effectively. The SAIPH algorithm used natural language processing to identify relevant tweets and then applied machine learning models to analyze these tweets for psychological markers. The data was processed using advanced text mining techniques to detect nuanced patterns indicative of suicidal ideation. The model achieved a high AUC of 0.88, indicating strong predictive power. This result underscores the algorithm's effectiveness in identifying individuals at risk based on their social media activity.

Future enhancements could include integrating more diverse data sources and improving the model's sensitivity to cultural and linguistic variations in expressing suicidal thoughts.

Moreover, Brown [58] investigated the use of Instagram data to predict acute suicidality. They employed both qualitative and quantitative language analyses and achieved promising results with machine learning models that detected suicidal ideation with high accuracy. Their findings underscore the importance of leveraging diverse social media platforms and sophisticated analytical techniques to enhance suicide prediction models. The study utilized a combination of NLP techniques to extract linguistic features from Instagram posts. Machine learning models were then trained to classify these features based on their association with suicidal ideation. The models showed high accuracy in detecting posts that indicated acute suicidal thoughts, demonstrating the utility of Instagram data in suicide prevention efforts. Expanding the dataset to include a broader range of social media platforms and integrating multimodal data (e.g., images and text) could further improve predictive accuracy.

Lastly, Bejan [59] highlighted the use of NLP to improve the ascertainment of suicidal ideation and suicide attempts in Electronic Health Records (EHRs). By employing information retrieval methodologies and a weakly supervised approach, they demonstrated that combining NLP with diagnostic codes significantly enhances the identification of suicidal behavior, providing a scalable and accurate tool for suicide prevention efforts. The study developed a scalable NLP approach to analyze over 200 million clinical notes from EHRs. Using a combination of word embeddings and information retrieval techniques, the researchers identified terms and phrases associated with suicidal ideation and attempts. The weakly supervised method involved ranking patients based on their relevance to suicide-related queries and validating the results through manual chart review. The NLP system showed high performance with an AUC of 98.6% for suicidal ideation and 97.3% for suicide attempts. The integration of diagnostic codes further enhanced the system's precision and recall. Future research could focus on refining the NLP algorithms to better handle negated expressions and explore the integration of real-time EHR data to provide timely interventions for at-risk patients.

Collectively, these studies underscore the potential of NLP and LLMs in leveraging social media data to predict suicidal behavior. The varied methodologies and datasets highlight the adaptability of these technologies to different contexts and their practical applications in mental health diagnostics. Future research should aim to refine these models further, improving accuracy and reducing the incidence of false positives and negatives, which are crucial for the reliable application of these technologies in mental health monitoring and intervention.

Study	***Models Used***	***Dataset and Data Collection***	***Key Focus***	***Results***	***Future Work***
Coppersmith [11]	Sentiment Analysis, LSTM, CNN	Twitter data	Identify linguistic markers of mental health issues	Effective in detecting mood states indicative of mental health concerns	None specified
Aldhyani [12]	CNN, BiLSTM, XGBoost	Reddit (SuicideWatch subreddit), 232,074 posts	Analyze patterns of suicidal ideation in Reddit posts	High accuracy in identifying posts with	None specified

				suicidal ideation	
Wang [33]	Various LLMs	Multiple social media platforms	Examine emotional and psychological expressions	Accurate prediction of suicidal tendencies	None specified
Fonseka [34]	Keyword extraction, thematic analysis	Clinical interviews and social media posts	Integrate clinical and social media data for mental health	Improved predictive accuracy for suicidal behaviors	None specified
Tadesse [35]	LSTM	Reddit (mental health-related subreddits)	Identify emotional and psychological states in subreddit posts	Effective identification of suicidal ideation	None specified
Roy [36]	NLP, machine learning	Twitter data	Predict future suicidal thoughts using Twitter data	AUC of 0.88 in predicting suicidal ideation	Integrate more diverse data sources and improve sensitivity to cultural and linguistic variations in expressing suicidal thoughts
Brown [37]	NLP, machine learning	Instagram data	Predict acute suicidality from Instagram posts	High accuracy in detecting posts indicating acute suicidal thoughts	Expand dataset to include more social media platforms and integrate multimodal data
Bejan [38]	NLP, word embeddings, information retrieval	Electronic Health Records (EHRs)	Improve ascertainment of suicidal ideation and attempts in EHRs	AUC of 98.6% for suicidal ideation, 97.3% for suicide attempts	Refine NLP algorithms to better handle negated expressions and explore real-time

					EHR data integration

Figure 10: Summary of research using NLP techniques for suicide prediction

2.7. Comparative Analysis of NLP models and LLMs for depression prediction

In the previous chapter we explored some research papers that have gone through comparative analysis of NLP models and LLMs, to result in important findings and other research papers comparing jonly LLMs.

Comparative analysis within the scope of NLP and LLMs provides an essential framework for improving depression prediction methodologies. By examining different computational models on standardized tasks and datasets, researchers can systematically identify the most effective approaches, tailor them to specific needs, and ensure robustness against diverse data sets. This analysis is crucial in mental health applications, where accurate detection and interpretation of depressive symptoms from text can significantly impact diagnosis and treatment plans [46].

The importance of comparative analysis in this field cannot be overstated. It not only sets benchmarks for what is achievable with current technology but also highlights the limitations and challenges of existing models. For instance, in the study "Comparative Analysis of NLP Models for Detecting Depression on Twitter" by Khush Gupta, Razaq Jinad, and Qingzhong Liu[60], several transformer-based models such as BERT, RoBERTa, DistilBERT, ALBERT, Electra, and XLNet were evaluated on their ability to detect depressive content from Twitter feeds. This study emphasized not just the accuracy but also the computational efficiency of these models, providing a nuanced view of their practicality in real-world applications.

The preprocessing of data in these studies involves several critical steps that directly affect the outcome of the models' performance. Techniques like tokenization, stop-word removal, and lemmatization are commonly employed to clean and standardize text data before it's fed into the models. Moreover, handling imbalanced datasets—common in depression detection where depressive instances may be significantly fewer than non-depressive ones—is another crucial aspect of the preprocessing phase that can heavily influence model accuracy.

Rashmi Rachh and Sanjana Kavatagi[61], evaluated the effectiveness of different embedding techniques including TF-IDF, Word2Vec, and BERT in detecting depressive content from social media platforms. This analysis utilized two distinct datasets to assess how these models performed across different contexts and label distributions. The first dataset, sourced from the Kaggle website, consisted of Twitter data categorized into depressive and non-depressive tweets. It included 2,242 tweets labeled as depressive and 2,502 tweets labeled as non-depressive, summing up to a total of 4,744 records.

The second dataset was derived from the shared task at RANLP 2023 titled "Detecting Signs of Depression from Social Media Text". This dataset provided a more granular categorization of depressive states, labeling entries as moderate, severe, or non-depressive. To streamline the analysis, moderate (3,678 records) and severe (768 records) labels, both indicative of depressive states, were combined into a single depressive category. The non-depressive category remained distinct, comprising 2,755 records. This adjustment led to a consolidated dataset of 7,201 records, allowing for a balanced comparison of depressive versus non-depressive content[33].

The use of these datasets enabled Rachh and Kavatagi to conduct a comprehensive evaluation of how different embedding techniques handle variably labeled data. Their findings highlighted that context-aware embeddings like BERT provided superior performance over traditional methods such as TF-IDF and Word2Vec. Specifically, BERT's ability to understand the semantic nuances of language proved effective in distinguishing between varying degrees of depressive content, showcasing its robustness across diverse datasets[33].

These datasets not only provided a platform to test the efficacy of embedding models but also contributed to understanding the scalability and adaptability of these models to different types of linguistic data and labeling schemes encountered in real-world applications[33].

The study by Kabir [4] used a dataset of 40,191 tweets labeled for depression severity to explore the efficacy of SVM, BiLSTM, BERT, and DistilBERT in classifying tweets into non-depressed and various depression severity levels based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). ROC curves from this study demonstrate the discriminative performance of each model, with BERT and DistilBERT showing high AUC values, indicating their effective differentiation between severity levels.

Zeberga [62] developed a sophisticated framework to identify depression and anxiety-related posts on social media, utilizing a blend of advanced NLP techniques and models. Their system integrated word2vec and BERT with a BiLSTM classifier, harnessing BERT's capability to capture the contextual and semantic nuances of text data related to mental health. The BiLSTM model was effectively used to process sequences, considering both preceding and succeeding contextual information in sentences. To enhance performance and efficiency, the study also applied knowledge distillation techniques, transferring insights from a large pretrained BERT model to a more compact model, DistilBERT. This approach allowed for refined model training and significant performance improvements. Data was systematically collected from social media platforms Reddit and Twitter, and through meticulous hyperparameter optimization, the model achieved an impressive accuracy of 98%, outperforming other models in similar applications. This study not only demonstrated the efficacy of combining multiple NLP techniques but also highlighted the potential of knowledge distillation in improving the performance of machine learning models in real-world applications.

The systematic review by Lin Gan, Yingqi Guo, and Tao Yang in "Machine Learning for Depression Detection on Web and Social Media" [20] further underscores the breadth of research in this area, summarizing various studies that have used machine learning techniques to detect depression. This review not only shows the diversity of approaches and models used but also highlights the emerging trend of using multimodal data (text, images, and user interaction data) to improve the accuracy of depression detection models.

In summary, the metrics used for comparative analysis in these studies typically include accuracy, precision, recall, and F1-score. Each metric offers insights into different aspects of model performance. Accuracy measures the overall correctness of the model across all classes, while precision and recall focus on the model’s performance in identifying true positive cases, which is critical in medical applications like depression detection where missing a positive case (low recall) or falsely identifying one (low precision) can have serious implications. The F1-score is particularly valuable as it provides a balance between precision and recall, offering a single measure of a model's accuracy at identifying true positives without being skewed by large numbers of true negatives.

In conclusion, comparative analysis in NLP and LLMs for depression prediction is a dynamic and evolving field that continuously adapts to the latest advancements in technology and methodology. The insights gained from these studies not only drive technological advancements but also enhance the practical applicability of these models in real-world scenarios, ultimately contributing to better mental health outcomes. The ongoing evolution of

model capabilities, coupled with the integration of multimodal data and sophisticated preprocessing techniques, remains crucial for further advancements in this vital field

The exploration of NLP models and LLMs for depression prediction opens up several avenues for future research that are crucial for enhancing their real-world applicability and accuracy, particularly in mental health diagnostics. Future studies could delve into optimizing computational resources, ensuring that advanced models are not only accurate but also efficient for real-time applications. This could include investigating more streamlined model architectures and engaging in techniques like knowledge distillation, which has demonstrated potential in maintaining performance while reducing complexity.

There is also significant potential in expanding the scope of data inputs by integrating multimodal data, including video, audio, and user interactions. This approach could provide a more comprehensive assessment of a user's mental health by capturing nuanced expressions of depression and anxiety that text alone might miss. Additionally, adapting these models to a broader array of social media platforms and online communities can offer insights across different cultural contexts and languages, enhancing the models' utility and impact.

Further investigations into context-aware embeddings, like BERT, which have shown superior performance in detecting depressive content, could focus on refining these technologies to better capture the complex and subtle language associated with mental health issues. This might involve enhancing the embeddings to more effectively interpret emotional nuances and idiomatic expressions that are typical in depressive dialogues.

As these models improve and potentially handle sensitive personal data, addressing ethical considerations and privacy becomes paramount. Future work should not only push the boundaries of technological capabilities but also develop frameworks and best practices that ensure the ethical use and deployment of these technologies. By focusing on these areas, researchers and practitioners can continue to improve the effectiveness and ethical integration of NLP and LLM technologies, ultimately contributing to better mental health outcomes and robust support systems.

Study	*Models Compared*	*Datasets and Social Media Platforms*	*Key Results*
Khushi Gupta, Razaq Jinad, and Qingzhong Liu [55]	BERT, RoBERTa, DistilBERT, ALBERT, Electra, XLNet	Twitter	Evaluated the ability to detect depressive content, emphasizing accuracy and computational efficiency.
Rashmi Rachh and Sanjana Kavatagi [61]	TF-IDF, Word2Vec, BERT	Kaggle (Twitter data), RANLP 2023 shared task	BERT outperformed traditional methods, effectively distinguishing varying degrees of depressive content across two datasets.

Kabir [4]	SVM, BiLSTM, BERT, DistilBERT	DEPTWEET dataset: 40,191 tweets labeled for depression severity	ROC curves demonstrated high AUC values for BERT and DistilBERT, effectively classifying tweets into varying depression levels.
Zeberga [62]	word2vec, BERT, BiLSTM	Reddit and Twitter	Achieved 98% accuracy with a knowledge distillation from BERT to DistilBERT, refining the model training significantly.

Figure 11: Summary of comparative analyses researches

Chapter 3: Methodology

3.1 Datasets Overview

In this study, three datasets from Kaggle are utilized to perform a comprehensive analysis of depression and suicide prediction using various NLP and LLM methods. Each dataset offers unique characteristics and insights into mental health discussions on different social media platforms. The selection of datasets for this study is crucial for several reasons. First and foremost, these datasets were chosen because they offer a diverse and comprehensive foundation for analyzing mental health issues using NLP and LLMs. The datasets include information organically written by individuals in various social contexts, making them highly representative of real-world mental health discussions.

One of the primary reasons for selecting these datasets is their suitability for binary classification tasks. Each dataset contains clear, binary labels that indicate the presence or absence of mental health-related content, such as depressive or suicidal thoughts. This binary classification is essential for training and evaluating predictive models that can accurately identify and categorize mental health issues.

Moreover, the 2 out of 3 selected datasets provide a balanced distribution of classes, ensuring that the models trained on them can learn effectively without being biased toward a particular class. This balance is critical for developing models that are both fair and accurate in their predictions.

Another significant factor in the dataset selection is the rich textual data they offer. The language used in these datasets is organic and reflects genuine conversations and expressions of mental health issues. This authenticity allows for a more realistic and practical analysis of how mental health is discussed in various social settings, making the findings more applicable and valuable.

Finally, these datasets also facilitate a robust comparison between NLP and LLM methods. By providing a common ground for evaluating different technological approaches, the datasets enable a comprehensive analysis of how well these models can understand and process natural language related to mental health. This comparison is vital for identifying the strengths and weaknesses of each method and for advancing the development of more effective mental health monitoring tools.

Reddit Depression Dataset

The Reddit Depression Dataset [63] consists of 7,650 posts from the r/SuicideWatch subreddit. This community is a support group where individuals share their experiences and seek support regarding suicidal thoughts and tendencies. The dataset includes the actual content of the posts, referred to as "clean_text," and binary labels indicating whether a post is related to suicidal ideation (1) or not (0). This labeling helps in distinguishing posts that are indicative of severe mental distress from general discussions. By analyzing these patterns and identifying linguistic cues associated with depression chance, this dataset provides a rich source of textual data indicative of severe mental health issues. The bar chart (Figure 12) shows the distribution of depression labels in the Reddit dataset. The labels are almost evenly distributed between "Not Depressed" (0) and "Depressed" (1). We can notice it is a balanced dataset.

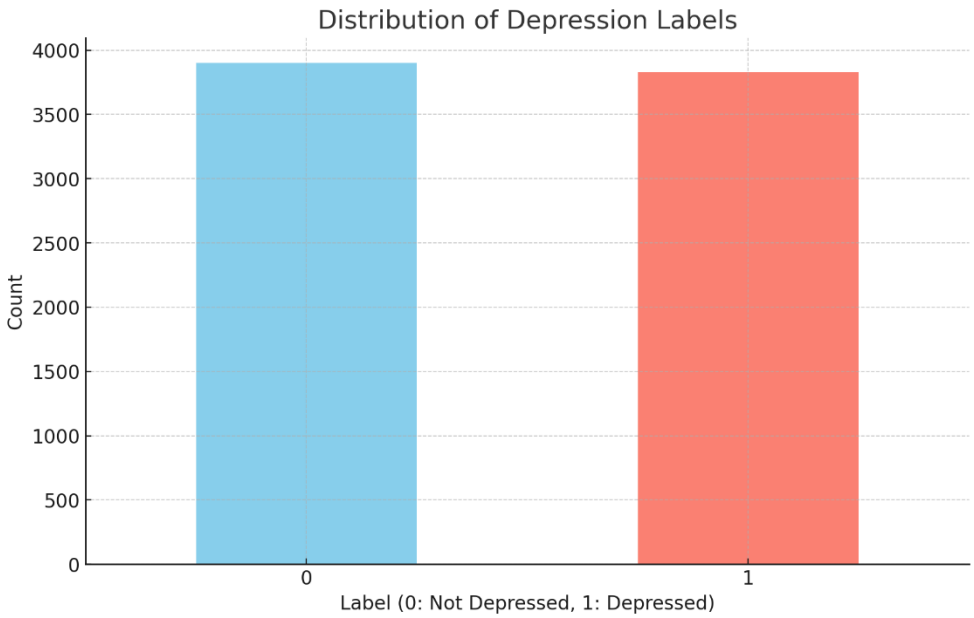

Figure 12:Distribution of Depression Labels in Reddit Dataset

Also, we can notice in the word cloud (Figure 13) below the most frequent words are found in the Reddit posts. Words like "life," "want," "know," "feel," "time," and "like" are prominently featured, indicating their common usage in the dataset.

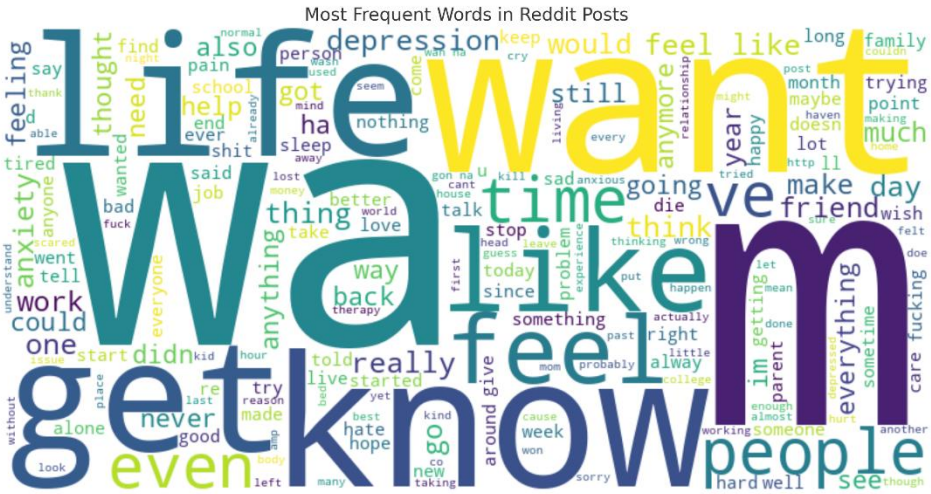

*Figure 13: Most frequent words in Reddit Depression Dataset *before preprocessing*

Additionally, the examination of bigrams, or pairs of consecutive words, sheds light on common themes and expressions within the Reddit posts. The top 20 most frequent bigrams include phrases like "don t," "feel like," "want to," and "going to," which hint at the context and narrative styles prevalent in the dataset. The recurrence of the word "depression" in multiple bigrams, such as "depression and," "depression is," and "my depression," emphasizes the central focus of the posts and the ways in which users discuss this topic.

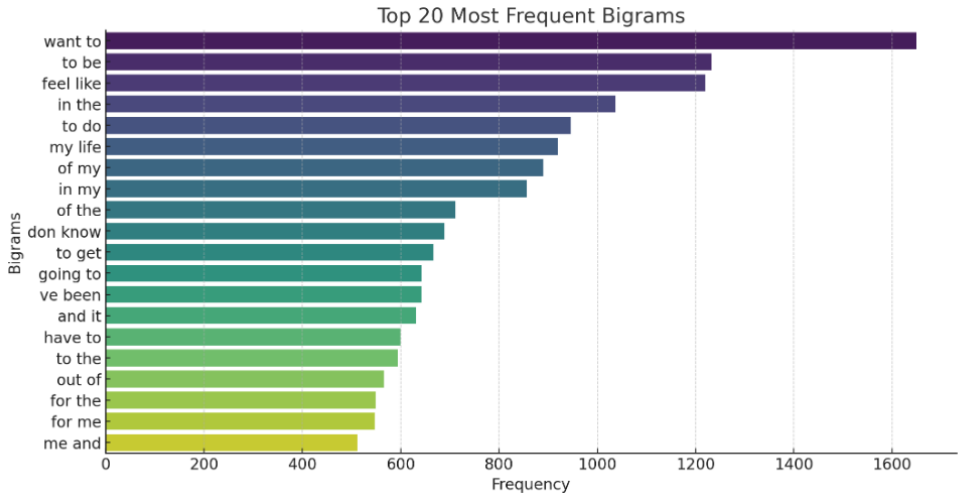

*Figure 14: Top 20 most frequent bigrams in Reddit Depression Dataset *before preprocessing*

Finally through the figure below, it is illustrated the average message length for posts labeled as depression-related (1) compared to those not related to depression (0). The significant difference in message lengths is noteworthy: depression-related posts have a considerably higher average length, exceeding 600 characters, whereas non-depression-related posts average around 100 characters. This discrepancy suggests that individuals discussing depression tend to write longer, more detailed posts, possibly to better articulate their feelings, experiences, and struggles. The lengthier posts may also indicate a greater need for expression and support, as users delve deeper into their narratives when sharing about depression.

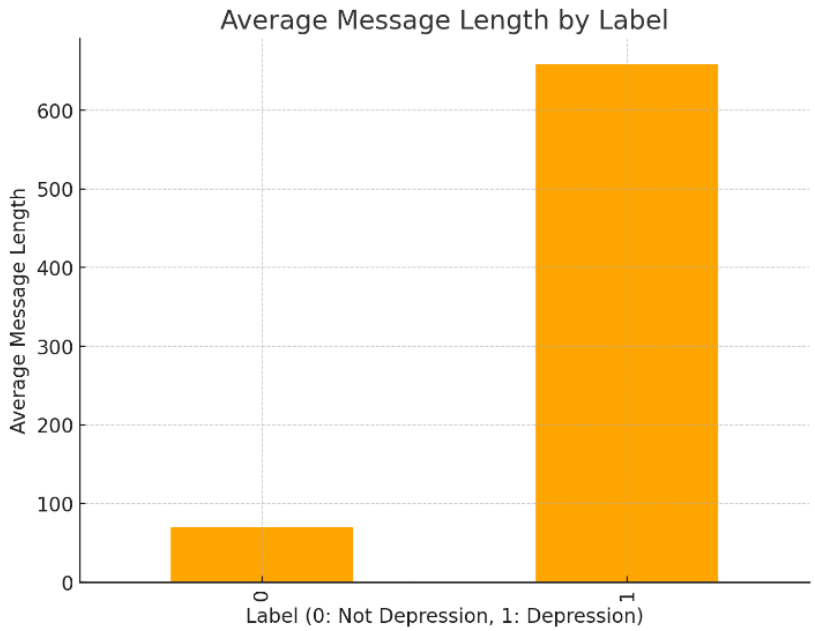

Figure 15: Average message length by label in Reddit Depression Dataset

Tweet Detection dataset

The Twitter Depression Dataset [64] comprises 10,282 tweets that express sentiments related to depression, anxiety, and stress. This dataset captures the public's real-time expressions of mental health issues on a widely used social media platform, providing a different perspective from the structured discussions in forums like Reddit. The dataset includes the actual tweet text, labeled as "message to examine," and binary labels indicating whether the tweet expresses depression (1) or not (0) (Figure 16). These labels help in identifying tweets that reflect depressive thoughts or emotions. Understanding how people publicly share their mental health experiences on Twitter offers valuable insights into the broader public discourse on mental health.

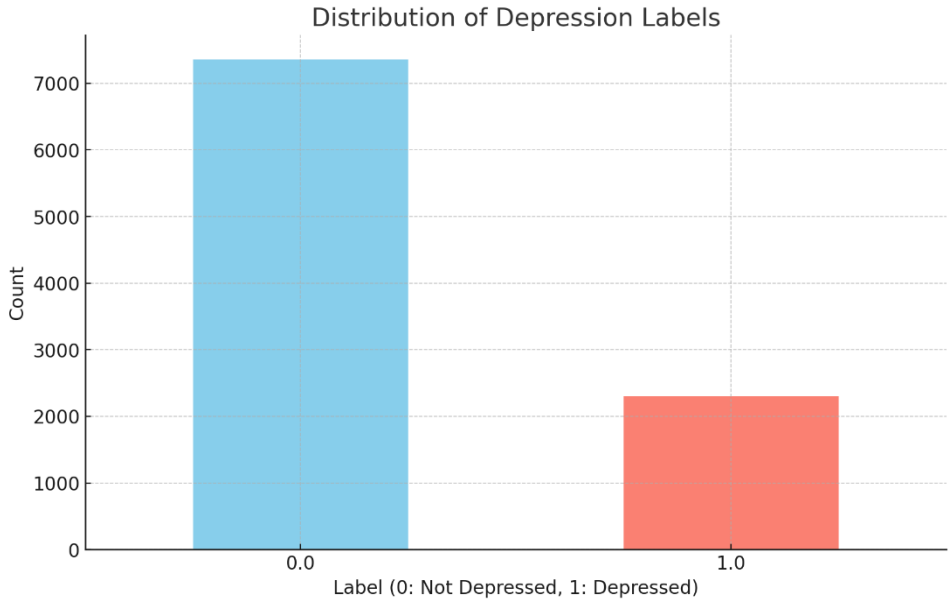

Figure 16: Distribution of Depression Labels

Also, we can notice in the word cloud (Figure 17) below the most frequent words are found in the tweets. Words like "depression," "love," "today," "thank," "time," and "good" are prominently featured, indicating their common usage in the dataset.

*Figure 17: Most frequent words in Twitter depression dataset *before preprocessing*

Additionally, the examination of bigrams, or pairs of consecutive words, sheds light on common themes and expressions within the tweets. The top 20 most frequent bigrams include phrases like "in the," "depression and," "for the," and "going to," which hint at the context and narrative styles prevalent in the dataset. The recurrence of the word "depression" in multiple bigrams emphasizes the central focus of the tweets and the ways in which users discuss this topic.

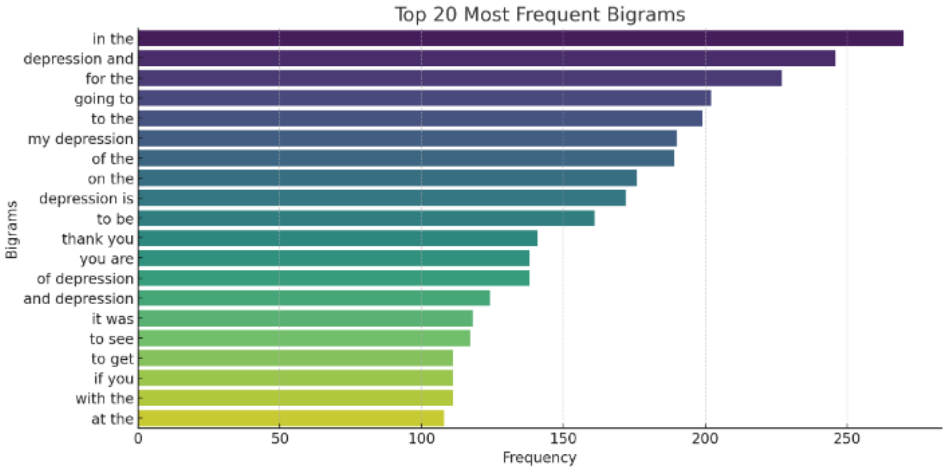

*Figure 18:Top 20 most frequent bigrams in Twitter Depression Dataset *before preprocessing*

In addition to the previously discussed insights, the Hashtag Analysis plot reveals the most frequently used hashtags in the dataset. Not surprisingly, hashtags like #depression, #anxiety, and #mentalhealth are among the most common, indicating the primary topics of concern in these tweets. Other notable hashtags include #therapy, #stress, and #mentalillness, highlighting related mental health issues.

These hashtags are crucial for understanding the broader context of the tweets, as they can help identify the specific aspects of mental health that users are discussing. This analysis

can also assist in tracking the spread of mental health awareness campaigns on social media and understanding the reach and impact of these discussions.

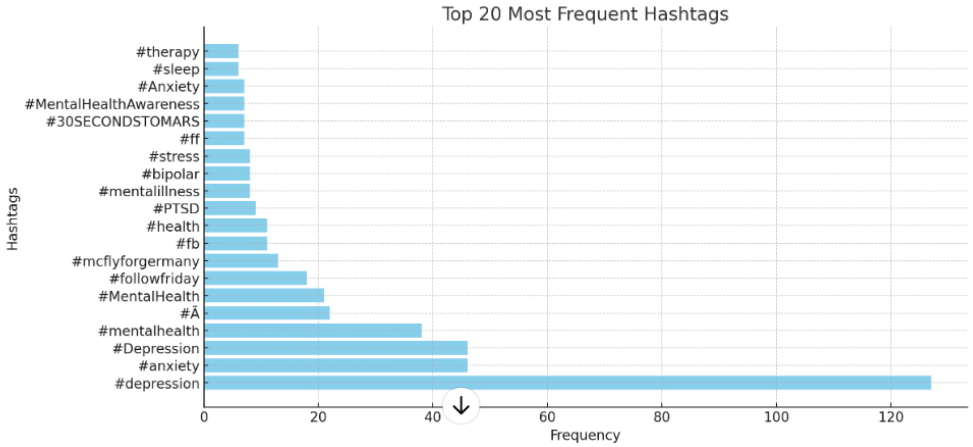

Figure 19: Top 20 most frequent hashtags in Twitter Depression Dataset before preprocessing

Finally, the analysis of the Twitter Depression Dataset further reveals intriguing patterns in the length and structure of tweets based on their sentiment labels. We can notice again, as shown in the bar chart depicting the average message length by label, tweets labeled as expressing depression (1) tend to be significantly longer than those not expressing depression (0). This suggests that individuals sharing depressive thoughts or emotions may use more words to articulate their feelings, possibly reflecting a need to elaborate on their experiences or to seek support from their audience.

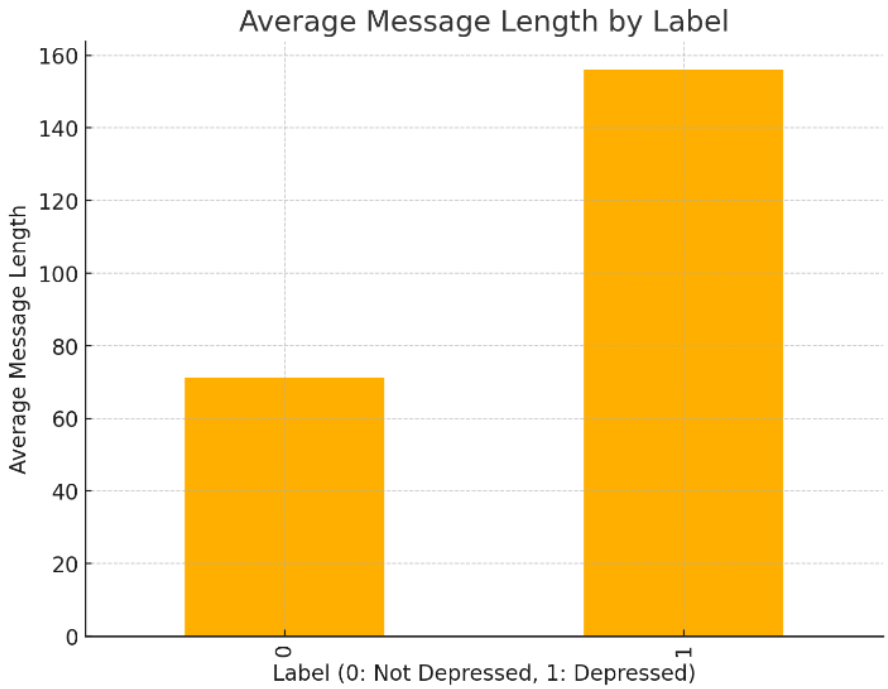

Figure 20: Average message length by label in Twitter Depression Dataset

Suicidal Tweet Detection

Finally we utilized the Suicidal Tweet Detection dataset [65] to train models for the automatic detection and flagging of tweets containing potential suicidal content. This enables platforms to take appropriate actions. The dataset includes a collection of tweets annotated to indicate whether each tweet is related to suicide. The primary goal is to develop and evaluate machine learning models that can classify tweets as expressing suicidal sentiments or not.

The dataset comprises two main columns: Tweet and Suicide. The Tweet column contains the text of tweets from various sources, covering diverse topics, emotions, and expressions. The Suicide column contains annotations classifying the tweets, with "Not Suicide post" for tweets that do not express suicidal sentiments, and "Potential Suicide post" for those that show indications of suicidal thoughts, feelings, or intentions.

The distribution of posts in the dataset is shown in the attached image. There are 1,102 tweets labeled as "Not Suicide post" and 659 tweets labeled as "Potential Suicide post". This visual representation highlights the dataset's composition and the balance between non-suicidal and potentially suicidal tweets (Figure 21).

This dataset is particularly useful for training and evaluating machine learning models to classify tweets as either non-suicidal or potentially suicidal. Researchers, data scientists, and developers can leverage this dataset to create systems that identify and flag concerning content on social media platforms, aiding early intervention and support for distressed individuals. The dataset can be employed for various NLP and sentiment analysis tasks, and we used it for suicide prediction.

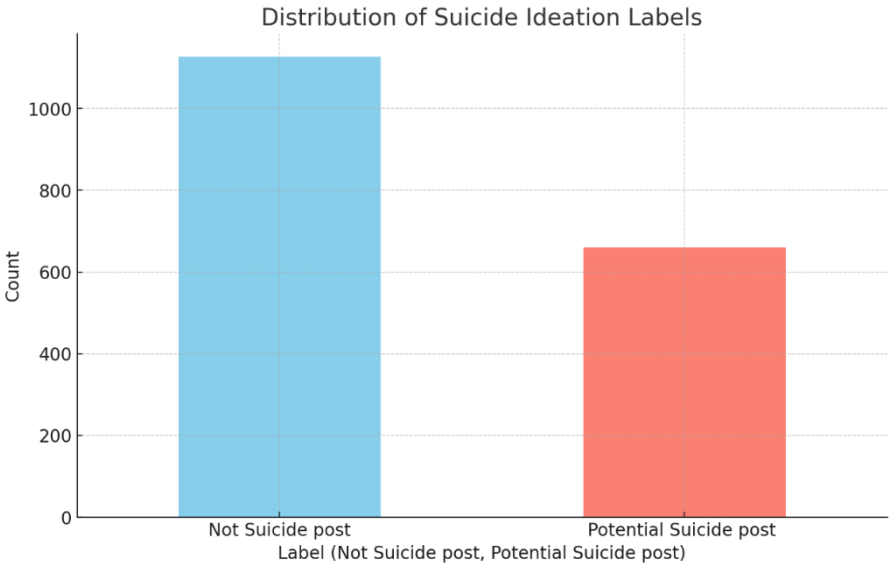

Figure 21: Distribution of Suicide Posts In Suicidal Tweet Detection dataset [65]

Also we have plotted the most frequent words in the dataset separately for suicide and not suicide posts. In tweets labeled as 'Potential Suicide posts,' the most frequent words often express distress and negative emotions, such as 'want,' 'don't,' 'hate,' 'life,' and 'tired,'

indicating potential signs of suicidal ideation, but we can also notice words like ‘https’, ‘co’, that will be cleared after preprocessing (Figure 22).

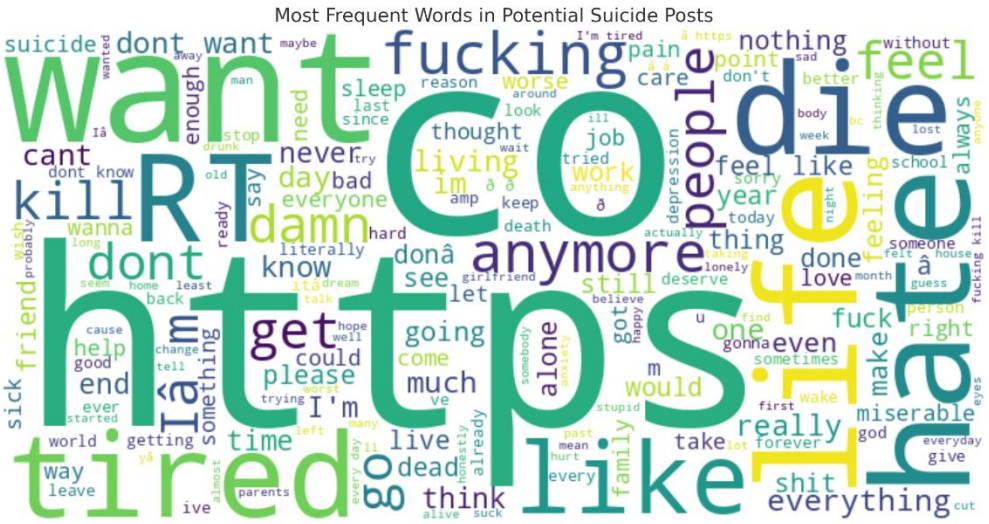

Figure 22: Most Frequent Words in Potential Suicide Posts

In tweets labeled as 'Not Suicide post,' the most common words include positive and neutral terms such as 'good,' 'want,' 'like,' 'love,' and 'happy,' reflecting everyday activities and emotions (Figure 23).

Figure 23: Most Frequent Words in Not Suicide Posts

Next, the figure below displays the top 20 most frequent bigrams found in the tweets. Bigrams are pairs of consecutive words that often appear together. The most common bigram is "https co," which likely represents URLs and can be disregarded for content analysis and they did through preprocessing. Other notable bigrams include "want to," "to be," "tired of," and "my life," indicating common expressions related to desires, existential reflections, and fatigue. The presence of phrases like "hate myself," "kill myself," and "to die" highlights the distressing nature of some of these tweets, reflecting severe emotional states and potential suicidal intentions. This analysis sheds light on the prevalent themes and expressions within the dataset, revealing the language patterns that characterize discussions of suicide ideation on Twitter.

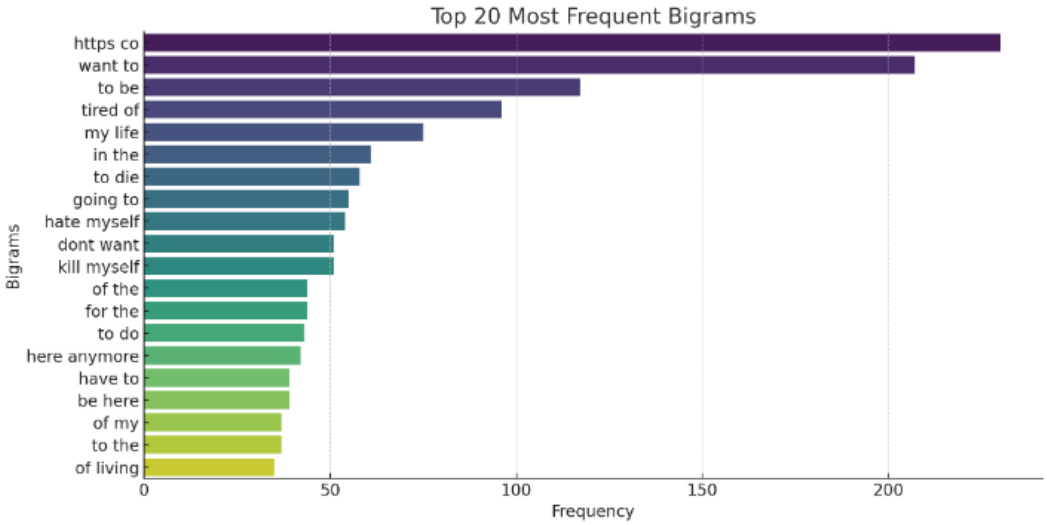

*Figure 24: Top 20 most frequent bigrams in Twitter Suicide Dataset *before preprocessing*

Lastly, the bar chart above illustrates the average message length for tweets labeled as "Not Suicide post" and "Potential Suicide post." As we have noticed in the analysis of depression-related content, individuals expressing potential suicide intentions also tend to use more words in their messages. The significant difference in message length is evident, with potential suicide posts having a much higher average length of around 160 characters, compared to approximately 80 characters for non-suicide posts. This pattern suggests that similar to individuals discussing depression, those expressing suicidal intentions may use longer messages to articulate their feelings and seek support, reflecting the complexity and urgency of their emotional state.

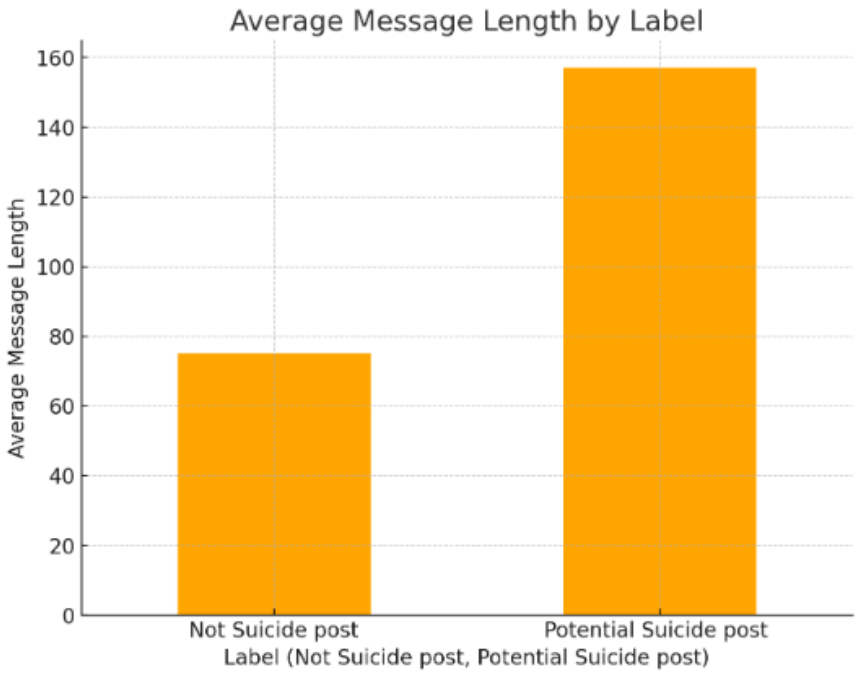

Figure 25: Average message length by label in Twitter Suicide Dataset

These three datasets collectively offer a diverse and comprehensive foundation for studying mental health issues using NLP and LLM methods. Each dataset brings unique perspectives from different platforms and types of discussions, enriching the analysis and

helping in developing robust predictive models for mental health conditions like depression and suicidal ideation. By leveraging the rich textual data and comprehensive labeling in these datasets, this study aims to advance the understanding and prediction of mental health issues in social media contexts.

In this research, we wanted to perform a comparative analysis for mental health prediction tasks. To achieve this, we followed a structured methodology outlined below.

3.2. Methodology

3.2.1 Datasets NLP methodology

In this chapter, we will analyze the NLP methodology used for the datasets. The methodology applied involves several key steps: data preprocessing, feature extraction, model training, and evaluation. This comprehensive approach ensures the development of robust predictive models for identifying depression-related content on social media.

Data preprocessing was the initial and essential step in preparing raw text data for analysis. It involved cleaning and transforming the text to remove noise and irrelevant information, thus improving the quality and effectiveness of subsequent NLP techniques.

Firstly, we removed HTML tags, which were often present in social media posts due to web formatting. These tags did not carry semantic information relevant to text analysis and were therefore removed [66].

Then, we converted all the text to lowercase. Lowercasing ensured that the model treated words like "Happy" and "happy" as the same word, which reduced the dimensionality of the data and helped create a more uniform dataset.

Next, we removed punctuation marks. Punctuation is generally not useful in understanding the sentiment or meaning of a text. Removing punctuation helped focus on the actual words and their meanings.

Following that, we performed tokenization, which is the process of splitting text into individual words or tokens. Tokenization was a fundamental step in NLP as it allowed the text to be analyzed on a word-by-word basis [67]

After tokenization, we removed stop words. Stop words are common words such as "and," "the,", and "is" that did not contribute significantly to the meaning of the text. Removing stop words helped in reducing noise and focusing on the important words.

Lastly, we applied lemmatization, which reduced words to their base or root form. For example, the words "running," "runs," and "ran" were reduced to "run". Lemmatization helped in treating different forms of a word as a single entity, thus reducing the complexity of the data.

Feature extraction transformed the cleaned text data into numerical representations that machine learning models could understand. This step was crucial for converting textual data into a format suitable for computational analysis.

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. It helped in identifying the most significant words in the text by balancing word frequency with inverse document frequency, highlighting words that were unique to specific [68]. The TF-IDF vectorizer was applied to the cleaned text data, transforming it into a matrix of TF-IDF features. This matrix represented the text in numerical form, capturing the importance of words relative to the entire dataset. These features were then used for model training and evaluation.

Word2Vec is a technique used to convert words into vector representations, capturing semantic relationships between them. It employed a neural network model to learn word associations from a large corpus of text. The resulting word vectors positioned words with similar meanings close to each other in the vector space. Word2Vec could be trained using two architectures: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicted the target word from the context words, while Skip-gram predicted the context words from the

target word[69]. Word2Vec embeddings were generated for the text data. This involved training the Word2Vec model on the cleaned text to produce dense vector representations for each word. These embeddings captured the semantic meanings of words, allowing for more nuanced text analysis. The resulting vectors were then used as features for model training and evaluation.

Model training involved applying machine learning algorithms to the extracted features to build predictive models. This step was critical for developing models that could accurately predict outcomes based on the input data. Several classifiers were used in this study, including Logistic Regression, and Naive Bayes. Each classifier has its strengths and is chosen based on the specific requirements of the task.

Logistic Regression is a widely used statistical model for binary classification problems. It estimates the probability that a given input belongs to a particular class by fitting a logistic function to the data. The model outputs probabilities that can be mapped to discrete classes using a threshold value. Logistic Regression is simple, interpretable, and effective for linearly separable data. Logistic Regression was applied to the TF-IDF and Word2Vec features. The model was trained to predict whether a post was indicative of depression based on the textual features. Hyperparameter tuning was performed to optimize the model's performance.

Naive Bayes is a probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions between features. It is particularly effective for text classification tasks due to its simplicity and efficiency. Despite its simplicity, Naive Bayes can often outperform more complex models, especially in high-dimensional spaces. The Naive Bayes classifier was trained using the TF-IDF and Word2Vec features. The model was evaluated for its accuracy and efficiency in classifying posts related to depression. The simplicity of Naive Bayes makes it a strong baseline for comparison with more complex models.

To sum up the prodedure above, we have created a flow diagram of our methodology as follows in the picture:

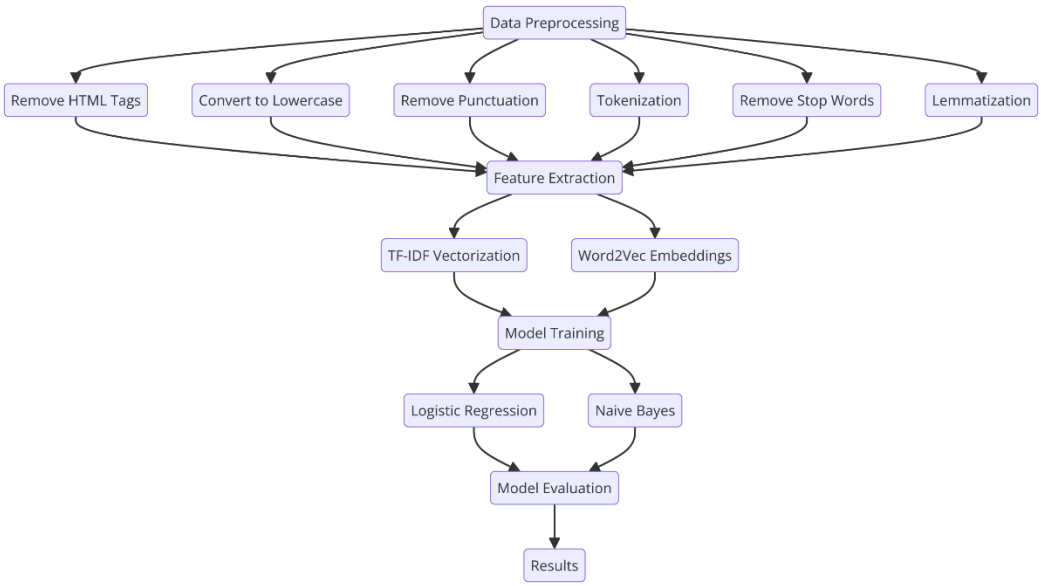

Figure 26: Prediction NLP procedure

3.2.1 BERT and Depression

For Twitter dataset we used the Kaggle notebook "Depressed Tweet Detection with DistilBERT" [70] for our analysis, which involves using the BERT model to detect depression in tweets.

The methodology for using BERT for depression detection on Twitter data involves several steps. The pre-trained BERT model was loaded using the transformers library. Specifically, the DistilBERT variant, which is a smaller, faster, and lighter version of BERT, was chosen. The BERT tokenizer was initialized to preprocess the text data. The tokenizer converts the tweets into token IDs that the BERT model can understand.

For data preprocessing, each tweet in the dataset was tokenized using the BERT tokenizer. This involved splitting the text into tokens, adding special tokens required by BERT, and padding or truncating the sequences to a fixed length. Attention masks were created to differentiate between actual tokens and padding tokens. These masks help BERT to focus on the real content of the tweets during training. Proper data preprocessing is crucial for improving model performance, as it helps in removing noise and making the data more suitable for training.

Fine-tuning the BERT model began with dataset preparation. The tokenized tweets, along with their attention masks and labels, were converted into a format suitable for the BERT model, typically using torch tensors. The pre-trained DistilBERT model was loaded, and the classification layer was initialized to adapt the model for the binary classification task (depressed vs. not depressed). The Trainer API from the transformers library was used to configure the training process. This included setting the learning rate, batch size, number of epochs, and other hyperparameters. The model was trained on the labeled tweet data. During training, the model's parameters were optimized to minimize the classification error on the training set.

After training, the model was tested on a separate test set to assess its performance. The test set consisted of tokenized tweets that the model had not seen during training. Various metrics such as accuracy, precision, recall, and F1-score were calculated to evaluate the model's performance, which are the same metrics used in all models we run. These metrics provided insights into how well the model could classify tweets as indicating depression or not. A high F1-score suggests a good trade-off between precision and recall, indicating the model's effectiveness.

The use of DistilBERT proves to be effective for sentiment analysis tasks, especially in detecting depressive tweets. Proper data preprocessing is crucial for improving model performance, as it helps in removing noise and making the data more suitable for training. Fine-tuning pre-trained models like DistilBERT on specific datasets can significantly enhance performance, leveraging the knowledge already embedded in these models. The attention masks help the model focus on the actual content, improving the training efficiency and accuracy.

Overall, the methodology outlined in the Kaggle notebook demonstrates a robust approach to using DistilBERT for detecting depressive tweets, highlighting the importance of each step from data preprocessing to model evaluation. The high performance metrics underscore the potential of transformer-based models in NLP tasks.

For depression detection on Reddit data, we used the methodology outlined in the Kaggle notebook "Fine-Tuning ROBERTA Base for Depression". This work follows similar steps to those used in the analysis of depressive tweets with DistilBERT but with some notable differences.

The dataset for this analysis is a cleaned version of a Reddit dataset containing posts related to depression. The data preprocessing steps involved removing noise such as special characters, stopwords, and unnecessary whitespace. The text was then tokenized using the ROBERTA tokenizer, converting the posts into token IDs that the model can understand.

Each Reddit post was tokenized using the ROBERTA tokenizer. This involved splitting the text into tokens, adding special tokens required by ROBERTA, and padding or truncating the sequences to a fixed length. Attention masks were created to differentiate between actual tokens and padding tokens, helping ROBERTA focus on the real content of the posts during

training. Proper data preprocessing is crucial for improving model performance, as it helps in removing noise and making the data more suitable for training.

Fine-tuning began with dataset preparation. The tokenized posts, along with their attention masks and labels, were converted into a format suitable for the ROBERTA model, typically using torch tensors. The pre-trained ROBERTA base model was loaded, and the classification layer was initialized to adapt the model for the binary classification task (depressed vs. not depressed). The Trainer API from the transformers library was used to configure the training process, including setting the learning rate, batch size, number of epochs, and other hyperparameters. The model was trained on the labeled Reddit data, and during training, the model's parameters were optimized to minimize the classification error on the training set.

The evaluation process was similar as before. The use of ROBERTA proves to be highly effective for sentiment analysis tasks, particularly in identifying depressive posts on Reddit. Proper data preprocessing significantly improves model performance by cleaning and structuring the input data. Fine-tuning pre-trained models like ROBERTA on specific datasets enhances performance by leveraging the pre-trained knowledge. Attention masks play a crucial role in helping the model focus on relevant parts of the text, improving training efficiency and prediction accuracy.

In summary, for depression detection on both Twitter and Reddit datasets, we followed a systematic approach involving data preprocessing, tokenization, and fine-tuning transformer-based models (DistilBERT and ROBERTA). The effectiveness of these models in sentiment analysis tasks underscores the importance of leveraging advanced NLP techniques and pre-trained models to achieve high performance in text classification tasks.

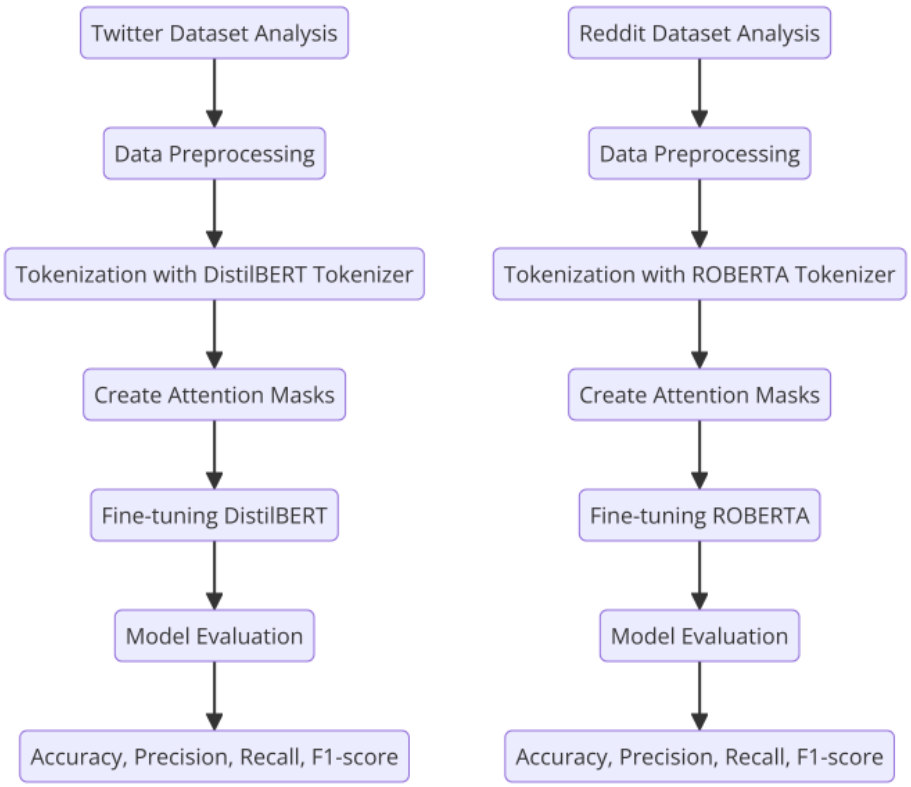

Figure 27: Depression prediction LLMs procedure

3.2.1 Suicide Prediction from Twitter

Lastly, the work from the Kaggle notebook "Suicide Tweet Detection Using BERT" [72] was employed to train models for the automatic detection and flagging of tweets containing

potential suicidal content. The primary objective was to develop and evaluate machine learning models capable of classifying tweets as expressing suicidal sentiments or not.

In this study, the dataset used comprised tweets labeled to indicate whether they are related to suicide. The initial steps in data preprocessing included loading the dataset and cleaning the text data by removing special characters, converting text to lowercase, and eliminating stopwords. Tokenization was applied to convert the text into sequences of words or tokens. Lemmatization was then used to reduce words to their base or root forms, improving the quality of the features extracted.

The BERT tokenizer was initialized to preprocess the text data, converting the tweets into token IDs that the BERT model can understand. The fine-tuning process began with dataset preparation. The tokenized tweets, along with their attention masks and labels, were converted into a format suitable for the BERT model, typically using torch tensors. The pre-trained BERT model was then fine-tuned on the labeled tweet data for the binary classification task (suicidal vs. non-suicidal).

The evaluation process was again similar, using metrics such as F1-score, Precision, Recall and Confusion Matrix.

The results of the model evaluation revealed high performance in classifying tweets related to suicide. The BERT model showed high accuracy, precision, recall, and F1-score, indicating its effectiveness in identifying tweets with potential suicidal content. The confusion matrix further demonstrated the model's capability in distinguishing between suicidal and non-suicidal tweets with minimal misclassifications.

The BERT model proved to be highly effective for the task of detecting suicidal tweets. Its ability to handle context and understand the intricacies of language contributed to its high performance. Proper text cleaning and lemmatization significantly improved the quality of the features, resulting in better model performance. The high scores in accuracy, precision, recall, and F1-score highlight the model's robustness and reliability in classifying suicidal tweets.

In summary, the analysis of tweets related to suicide using the BERT model unveiled crucial insights into the nature of these posts. The methodology and results underscored the effectiveness of advanced NLP techniques and machine learning models in analyzing and classifying text data related to suicidality, highlighting the importance of proper preprocessing and feature extraction.

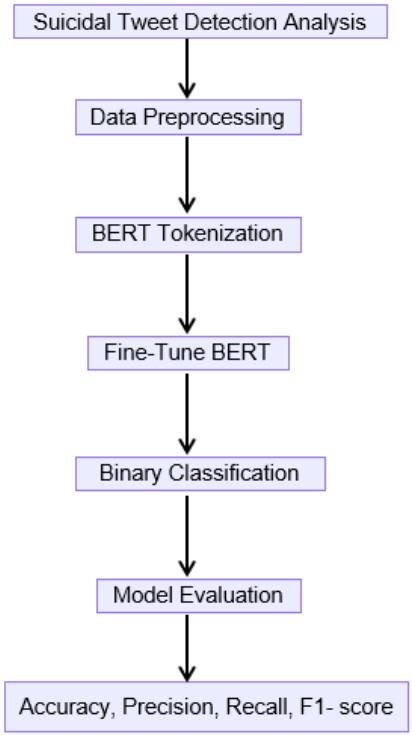

Figure 28: Suicide prediction Bert procedure

Chapter 4: Results

In this chapter, the results of the analysis on the three distinct datasets are presented, i.e., the Depression Twitter Dataset, the Reddit Depression Dataset, and the Suicide Twitter Dataset. Various machine learning models and techniques were employed to classify the data, with a focus on evaluating the performance of each model using the classification report and confusion matrix. The F1-score and accuracy were used as the final metrics for the analysis, providing a comprehensive understanding of each model's effectiveness.

For the Depression Twitter Dataset, four different models were explored: Logistic Regression with TF-IDF vectorization, Naive Bayes with TF-IDF vectorization, Logistic Regression with Word2Vec embeddings, and DistilBERT. Each model was trained and tested on the dataset, and their performance was evaluated using classification reports and confusion matrices. The results provided insights into the strengths and weaknesses of each approach in detecting depression-related content in tweets.

The Reddit Depression Dataset was analyzed using a similar methodology. The models used were Logistic Regression with TF-IDF vectorization, Naive Bayes with TF-IDF vectorization, Logistic Regression with Word2Vec embeddings, and RoBERTa. By applying these models to the Reddit dataset, an understanding was aimed at regarding how well each model could classify posts related to depression. The evaluation was based on classification reports and confusion matrices, with the F1-score and accuracy as the key metrics.

For the Suicide Twitter Dataset, two models were employed: Random Forest with TF-IDF vectorization and BERT. These models were evaluated to determine their effectiveness in identifying suicidal content in tweets. The analysis highlighted the F1-score and accuracy of each model.

Through this systematic evaluation, the aim is to identify the most effective models and techniques for each dataset, providing a robust framework for future research and practical applications in the field of mental health detection using social media data.

Depression Reddit Dataset

Logistic Regression TFIDF

The performance of the classification model for Logistic Regression TD-IDF on Depression Reddit dataset, evaluated using several metrics, is summarized in the classification report. The report indicates that for Class 0 (Non-depression), the model achieved a precision of 0.93, a recall of 0.98, and an F1-score of 0.96, with support for 390 instances. For Class 1 (Depression), the precision was 0.98, the recall was 0.93, and the F1-score was 0.95, with support for 383 instances. The overall performance of the model is reflected in an accuracy of 0.95. Additionally, the macro and weighted averages for precision, recall, and F1-score are all 0.96 and 0.95 respectively. These results demonstrate that the model performs well, with a high level of precision and recall for both classes, indicating strong performance in classifying instances of depression and non-depression.

	precision	recall	f1-score	support
0	0.93	0.98	0.96	390
1	0.98	0.93	0.95	383
accuracy			0.95	773
macro avg	0.96	0.95	0.95	773
weighted avg	0.96	0.95	0.95	773

Figure 29: Classification report for Logistic Regression TD-IDF on Depression Reddit Dataset

The confusion matrix provides a visual representation of the model's performance in distinguishing between the two classes. The matrix shows that the model correctly classified the majority of instances in both classes, with 382 true positives and 356 true negatives. There are relatively low numbers of false positives (27) and false negatives (8). This indicates that the model is effective in accurately identifying both depressive and non-depressive posts.

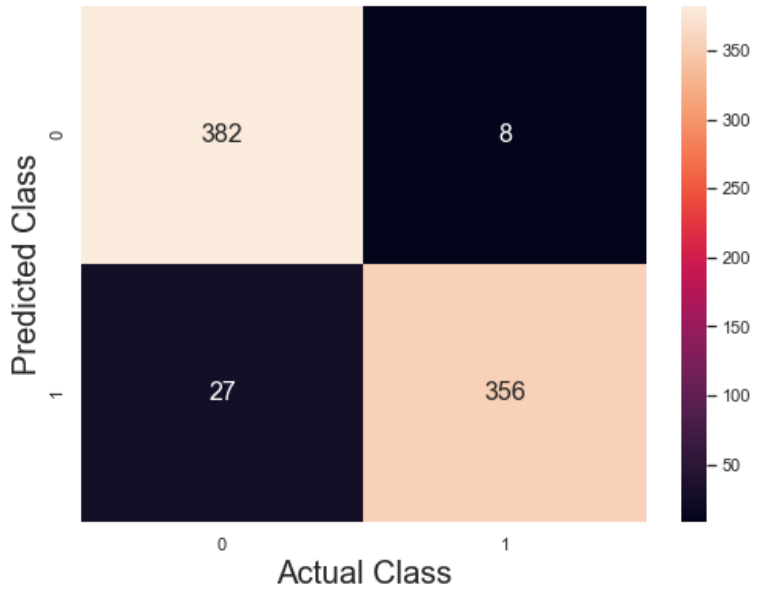

Figure 30: Confusion matrix for Logistic Regression TD-IDF on Depression Reddit Dataset

The model's interpretability was enhanced by analyzing the feature importance, which indicates the most significant features in distinguishing between depression and non-depression. The top ten positive features indicative of depression include terms such as "depression," "anxiety," "life," "pression," "feel," "fuck," "help," and "anymore." Conversely, the top ten negative features indicative of non-depression include terms such as "update," "omg," "haha," "cold," "poor," "ugh," "sick," "sad," "oh," and "miss." These features were identified using the Random Forest TF-IDF method, and their coefficient values highlight their relative importance in the classification process. The term "depression" has the highest positive coefficient, indicating its strong association with depressive posts, while terms like "update" and "omg" have the most significant negative coefficients, suggesting their association with non-depressive content.

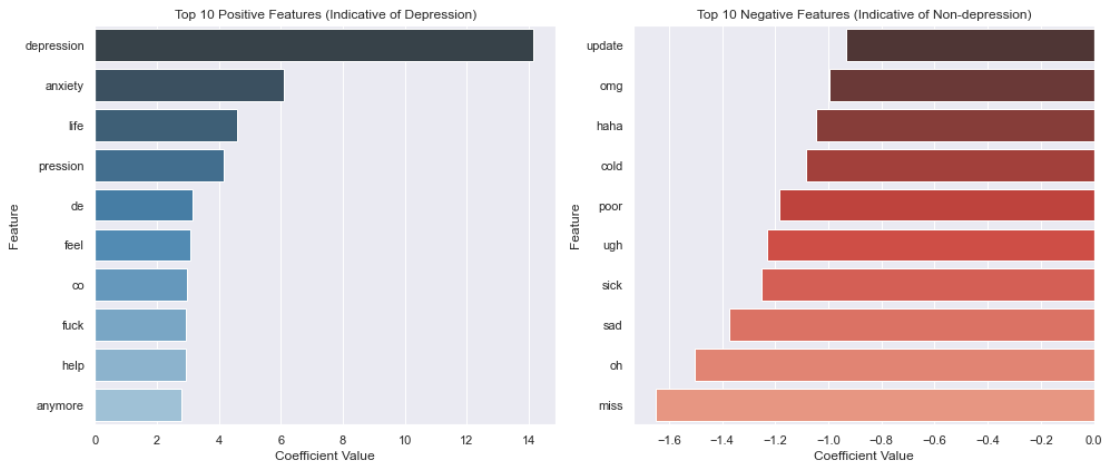

Figure 31: Top 10 positive and negative features for Logistic Regression TD-IDF on the Depression Reddit Dataset

Naïve Bayes TFIDF

The performance of the classification model, evaluated using several metrics, is summarized in the classification report. For Class 0 (Non-depression), the model achieved a

precision of 0.95, a recall of 0.81, and an F1-score of 0.87, with support for 390 instances. For Class 1 (Depression), the precision was 0.83, the recall was 0.96, and the F1-score was 0.89, with support for 383 cases. The overall accuracy of the model is 0.88. Additionally, the macro and weighted averages for precision, recall, and F1-score are all 0.89 and 0.88 respectively. These results indicate a reasonably good performance of the model, with a notable ability to recall depressive posts.

	precision	recall	f1-score	support
0	0.95	0.81	0.87	390
1	0.83	0.96	0.89	383
accuracy			0.88	773
macro avg	0.89	0.88	0.88	773
weighted avg	0.89	0.88	0.88	773

Figure 32: Classification report for Naïve Bayes TD-IDF on Depression Reddit Dataset

The confusion matrix provides a visual representation of the model's performance in distinguishing between the two classes. The matrix shows that the model correctly classified 314 instances of non-depression and 366 instances of depression. There are 76 false positives and 17 false negatives. This indicates that while the model is effective in identifying depressive posts, it has a higher rate of false positives.

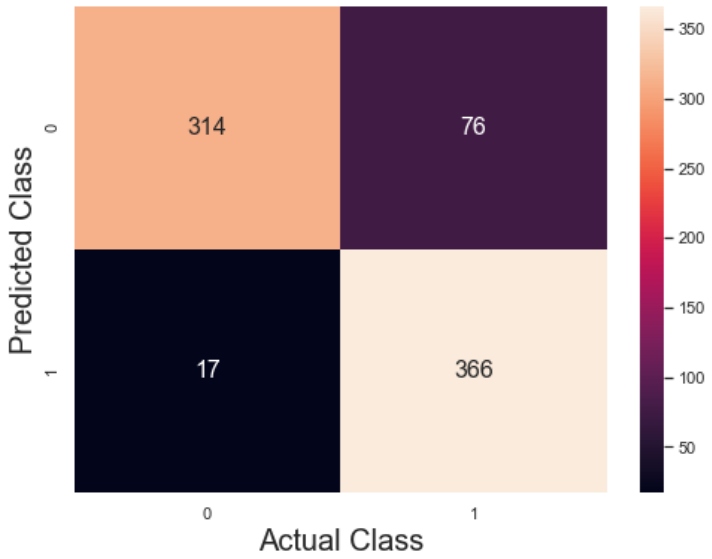

Figure 33: Confusion matrix for Naïve Bayes TD-IDF on Depression Twitter Dataset

The model's interpretability was enhanced by analyzing the feature importance, which indicates the most significant features in distinguishing between depression and non-depression. The top ten positive features indicative of depression include terms such as "depression," "feel," "like," "get," "want," "go," "know," "anxiety," and "life." Conversely, the top ten negative features indicative of non-depression include terms such as "haha," "window," "bar," "omg," "holiday," "ugh," "www," "homework," "update," and "rain." These

features were identified using the Logistic Regression TFIDF method, and their coefficient values highlight their relative importance in the classification process. The term "depression" has the highest positive coefficient, indicating its strong association with depressive posts, while terms like "haha" and "window" have the most significant negative coefficients, suggesting their association with non-depressive content.

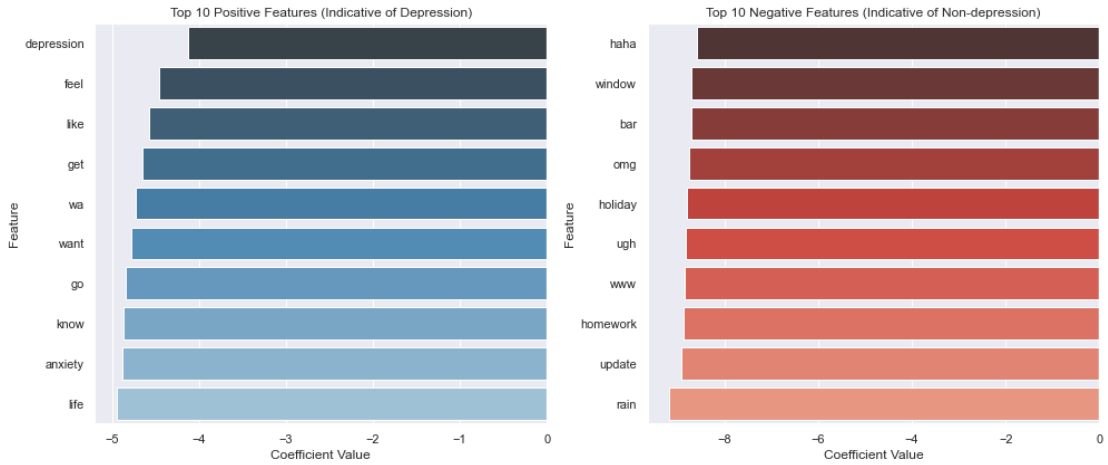

Figure 34: Top 10 positive and negative features for Naïve Bayes TD-IDF on the Depression Reddit Dataset

Logistic Regresion W2Vec

The performance of the classification model, evaluated using several metrics, is summarized in the classification report. For Class 0 (Non-depression), the model achieved a precision of 0.83, a recall of 0.81, and an F1-score of 0.82, with support for 390 instances. For Class 1 (Depression), the precision was 0.81, the recall was 0.83, and the F1-score was 0.82, with support for 383 instances. The overall accuracy of the model is 0.82. Additionally, the macro and weighted averages for precision, recall, and F1-score are all 0.82. These results indicate a balanced performance of the model, with both classes achieving similar levels of precision, recall, and F1-scores.

	precision	recall	f1-score	support
0	0.83	0.81	0.82	390
1	0.81	0.83	0.82	383
accuracy			0.82	773
macro avg	0.82	0.82	0.82	773
weighted avg	0.82	0.82	0.82	773

Figure 35: Classification report for Logistic Regression W2Vec on Depression Reddit Dataset

The confusion matrix provides a visual representation of the model's performance in distinguishing between the two classes. The matrix shows that the model correctly classified 314 instances of non-depression and 318 instances of depression. There are 76 false positives and 65 false negatives. This indicates a moderate level of effectiveness in distinguishing between depressive and non-depressive posts, with a somewhat higher rate of misclassification compared to the other models.

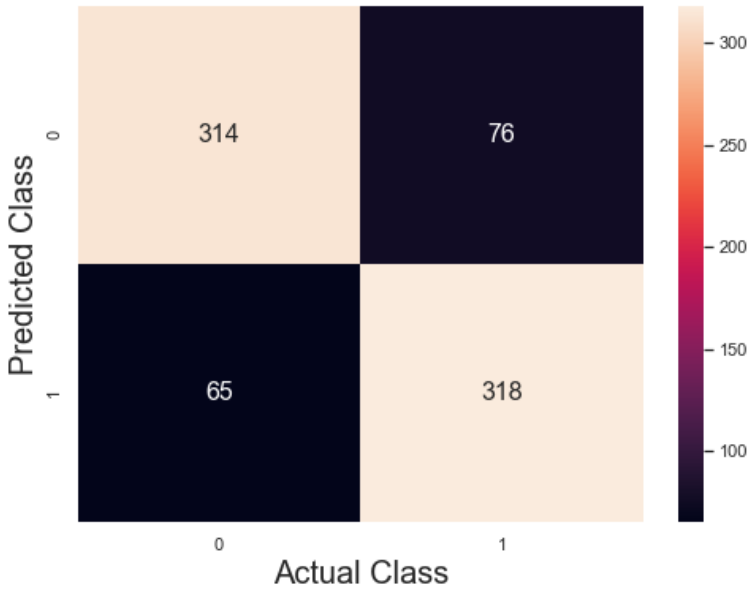

Figure 36: Confusion matrix for Logistic Regression W2Vec on Depression Reddit Dataset

Roberta Depression Reddit

The training performance of the model, specifically using the Roberta model for identifying depression on Reddit, is summarized across three epochs. The training loss, validation loss, accuracy, and F1 score are tracked for each epoch:

In the first epoch, the training loss was 0.218600, the validation loss was 0.086172, the accuracy was 0.958096, and the F1-score was 0.958053.

In the second epoch, the training loss decreased to 0.125800, the validation loss increased slightly to 0.089713, the accuracy improved to 0.981376, and the F1-score increased to 0.980624.

In the third epoch, the training loss further decreased to 0.067400, the validation loss decreased to 0.066574, the accuracy improved to 0.983445, and the F1-score increased to 0.982906.

These results indicate that the Roberta model exhibited strong performance improvements over the training epochs, culminating in an accuracy of 98.29%, an F1-score of 0.98, and a validation loss of 0.067 in the final epoch.

In summary, the results indicate that the Roberta model is highly effective in identifying depression-related posts on Reddit, as reflected in the high accuracy and F1 score. The classification metrics, confusion matrix, and training performance collectively demonstrate the model's robust performance.

Comparative Analysis Depression Reddit Dataset

The comparative analysis of different models used for classifying depression-related posts on the Reddit dataset reveals distinct performance levels in terms of F1-scores and accuracy rates. The Roberta model emerges as the top performer with an F1-score of 0.98 and an accuracy of 98.29%, showcasing its exceptional capability in handling natural language data. The Random Forest TFIDF model also shows robust results with both F1-score and accuracy at 0.95. The Logistic Regression TFIDF model demonstrates strong performance with an F1-score of 0.95 and an accuracy of 0.95. In contrast, the Naïve Bayes TFIDF model, with an F1-score of 0.89 and an accuracy of 0.88, shows good effectiveness but is slightly outperformed by the

other models. Lastly, the Logistic Regression W2Vec model displays relatively lower performance with an F1-score of 0.82 and an accuracy of 0.82.

Model	F1-Score	Accuracy
Logistic Regression TFIDF	0.95	0.95
Naïve Bayes TFIDF	0.89	0.88
Logistic Regression W2Vec	0.82	0.82
Roberta	0.98	0.98

Figure 37: Comparative analysis table for NLP models in Depression Reddit Dataset

Depression Twitter

Logistic Regression TFIDF

The classification report for the Logistic Regression model using TF-IDF vectorization on the Depression Twitter Dataset shows high performance with an accuracy of 0.99. The precision, recall, and F1-score for class 0.0 (non-depressed) are 0.99, 1.00, and 0.99, respectively. For class 1.0 (depressed), the precision is 1.00, recall is 0.95, and the F1-score is 0.98. The macro average for precision, recall, and F1-score are 0.99, 0.98, and 0.98, respectively, while the weighted average for all metrics is 0.99.

	precision	recall	f1-score	support
0.0	0.99	1.0	0.99	730
1.0	1.0	0.95	0.98	232
accuracy			0.99	962
macro avg	0.99	0.98	0.98	962
weighted avg	0.99	0.99	0.99	962

Figure 38: Classification report for Logistic Regression TD-IDF on Depression Twitter Dataset

The confusion matrix was also generated, highlighting the model's effectiveness. It shows 730 true positives and 221 true negatives, with 11 false negatives and no false positives. This indicates that the model is highly accurate in distinguishing between depressed and non-depressed tweets.

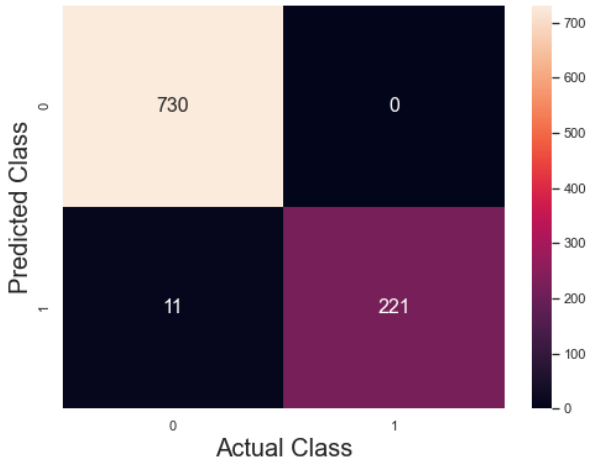

Figure 39: Confusion matrix for Logistic Regression TD-IDF on Depression Twitter Dataset

The feature importance graph was produced, displaying the top 10 positive and negative features indicative of depression and non-depression. Terms like "depression," "anxiety," "emoji," and "mental" are strong indicators of depression, while words such as "movie," "tomorrow," "love," and "thank" are indicative of non-depression. This visualization provides insights into the key features that the model uses to make its predictions. This is a good indicator that the model performs with good understanding of the task.

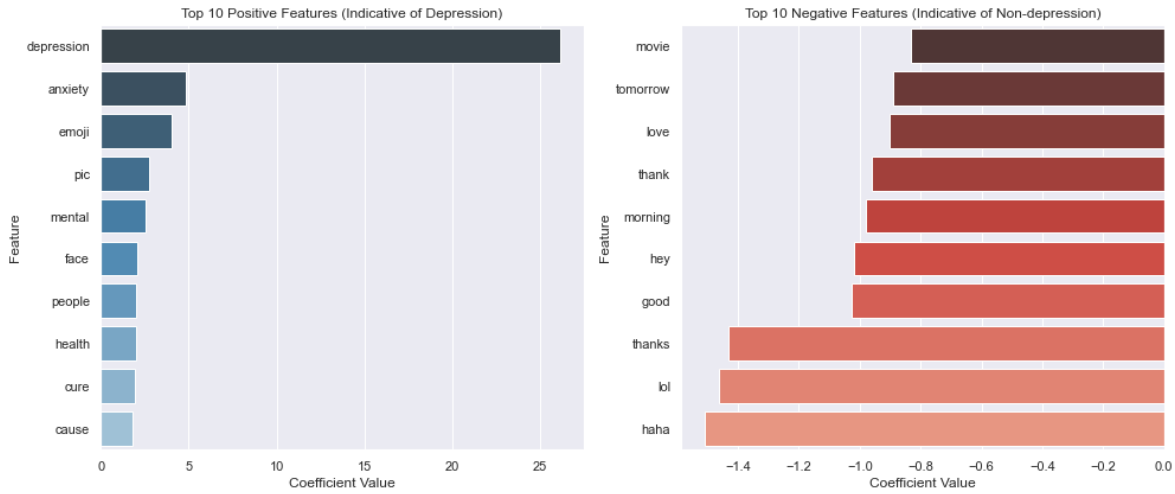

Figure 40: Top 10 positive and negative features for Logistic Regression TD-IDF on the Depression Twitter Dataset

Naïve Bayes TFIDF

The classification report for the Naive Bayes model using TF-IDF vectorization on the Depression Twitter Dataset shows high performance with an accuracy of 0.96. The precision, recall, and F1-score for class 0.0 (non-depressed) are 0.96, 0.99, and 0.97, respectively. For class 1.0 (depressed), the precision is 0.95, recall is 0.86, and the F1-score is 0.90. The macro average for precision, recall, and F1-score are 0.95, 0.92, and 0.94, respectively, while the weighted average for all metrics is 0.96.

	precision	recall	f1-score	support
0.0	0.96	0.99	0.97	730
1.0	0.95	0.86	0.9	232
accuracy			0.96	962
macro avg	0.95	0.92	0.94	962
weighted avg	0.96	0.96	0.96	962

Figure 41: Classification report for Naïve Bayes TD-IDF on Depression Twitter Dataset

The confusion matrix was also generated, highlighting the model's effectiveness. It shows 720 true positives and 200 true negatives, with 10 false negatives and 32 false positives. This indicates that the model is highly accurate in distinguishing between depressed and non-depressed tweets.

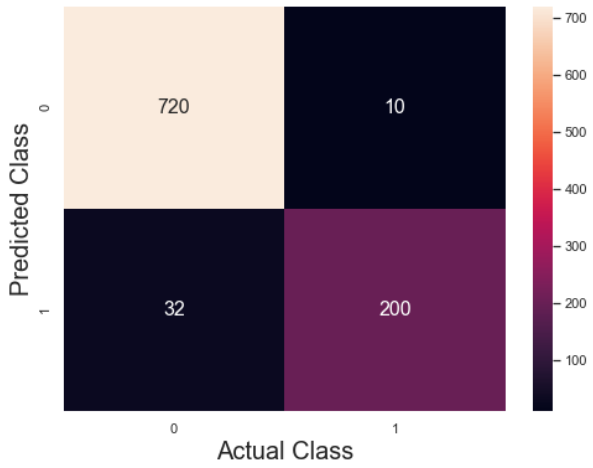

Figure 42: Confusion matrix for Naïve Bayes TD-IDF on Depression Twitter Dataset

The feature importance graph was produced, displaying the top 10 positive and negative features indicative of depression and non-depression. Terms like "depression," "anxiety," "pic," and "emoji" are strong indicators of depression, while words such as "bday," "bbq," "bb," and "bar" are indicative of non-depression. This visualization provides insights into the key features that the model uses to make its predictions.

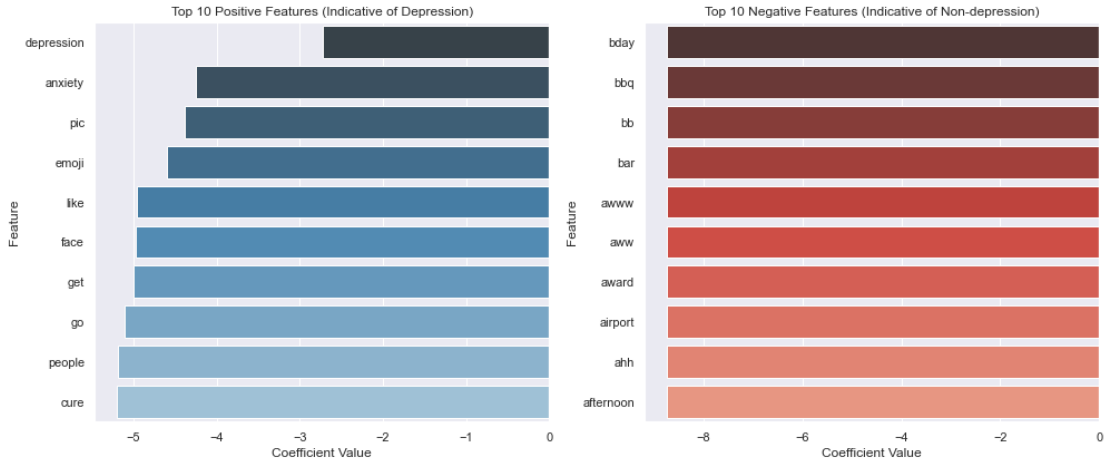

Figure 43: Top 10 positive and negative features for Naïve Bayes TD-IDF on the Depression Twitter Dataset

Logistic Regression W2Vec

The classification report for the Logistic Regression model using Word2Vec embeddings on the Depression Twitter Dataset shows an overall accuracy of 0.87. The precision, recall, and F1-score for class 0.0 (non-depressed) are 0.85, 1.00, and 0.92, respectively. For class 1 (depressed), the precision is 0.98, recall is 0.46, and the F1-score is 0.63. The macro average for precision, recall, and F1-score are 0.92, 0.73, and 0.77, respectively, while the weighted average for all metrics is 0.88.

	precision	recall	f1-score	support
0.0	0.85	1.0	0.92	730
1.0	0.98	0.46	0.63	232
accuracy			0.87	962
macro avg	0.92	0.73	0.77	962
weighted avg	0.88	0.87	0.85	962

Figure 44: Classification report for Logistic Regression W2Vec on Depression Twitter Dataset

The confusion matrix highlights the model's performance, showing 728 true positives and 107 true negatives. There are 125 false positives and 2 false negatives, indicating a lower accuracy in distinguishing between depressed and non-depressed tweets compared to other models.

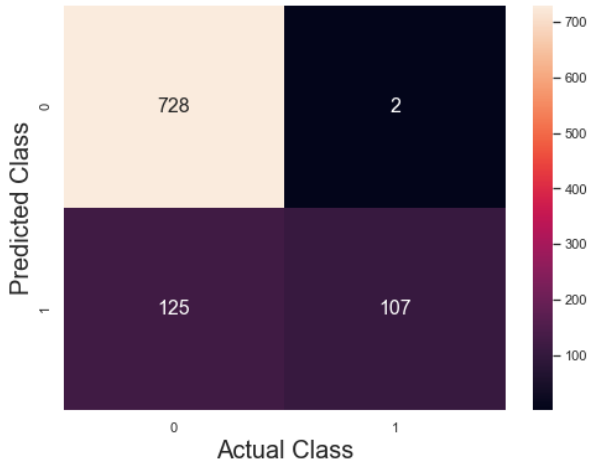

Figure 45: Confusion Matrix for Logistic Regression W2Vec on Depression Twitter Dataset

DistilBERT Depression Twitter

The classification report for the DistilBERT model on the Depression Twitter Dataset shows exceptional performance with an overall accuracy of 99.9%. The precision, recall, and F1-score for the "Not Depressed" class are 0.99, 1, and 0.99, respectively. For the "Depressed" class, the precision is 1, recall is 0.99, and the F1-score is 0.99. The macro average for precision, recall, and F1-score are 0.99, 0.99, and 0.99, respectively, while the weighted average for all metrics is 0.99. A high F1-score suggests a good trade-off between precision and recall, indicating the model's effectiveness.

	precision	recall	f1-score	support
Not Depressed	0.9988	1.0	0.9994	1600
Depressed	1.0	0.9956	0.9978	457
accuracy			0.999	2057
macro avg	0.9994	0.9978	0.9986	2057
weighted avg	0.999	0.999	0.999	2057

Figure 46: Classification report for the DistilBERT model on the Depression Twitter Dataset

The confusion matrix further confirms the model's effectiveness, with 1600 true positives and 457 true negatives. There are 0 false negatives and 4 false positives, indicating the model's high accuracy in distinguishing between depressed and non-depressed tweets.

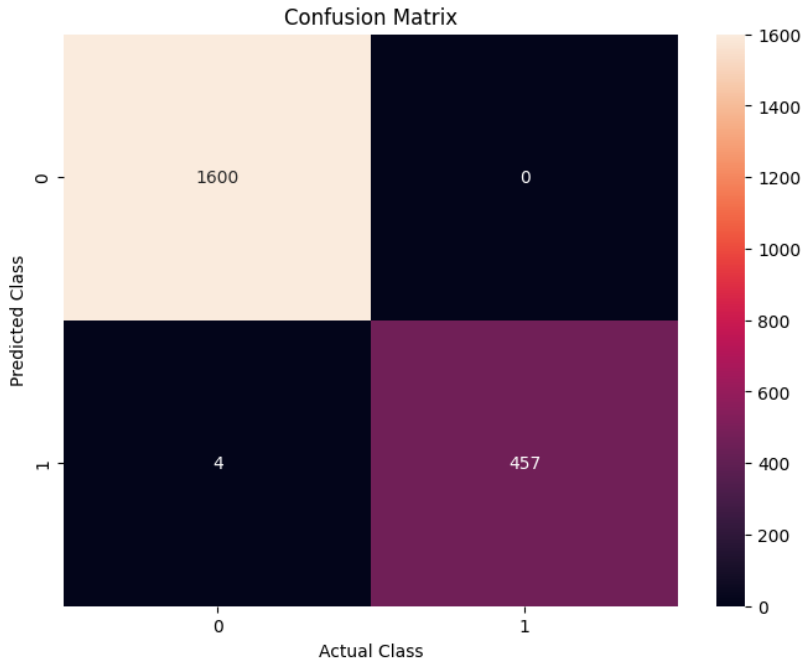

Figure 47: Confusion matrix for the DistilBERT model on the Depression Twitter Dataset

Comparative analysis Depression Twitter Dataset

The comparative analysis of various models applied to the Depression Twitter Dataset reveals distinct performance levels in terms of F1-scores and accuracy rates. DistilBERT stands out with the highest F1-score of 0.99 and an impressive accuracy of 0.999, indicating its superior capability in accurately identifying both depressed and non-depressed tweets. The Logistic Regression TFIDF model also demonstrates strong performance with an F1-score of 0.98 and an accuracy of 0.99, reflecting an effective balance between precision and recall. In contrast, the Naïve Bayes TFIDF model, with an F1-score of 0.90 and an accuracy of 0.96, shows good effectiveness but is slightly outperformed by DistilBERT and Logistic Regression TFIDF. Lastly, the Logistic Regression W2Vec model displays relatively lower performance with an F1-score of 0.63 and an accuracy of 0.87. These results underscore the high capability of advanced models like DistilBERT and Logistic Regression TFIDF in handling complex text classification tasks related to mental health on social media platforms.These results underscore the high capability of advanced models like DistilBERT and Logistic Regression TFIDF in handling complex text classification tasks related to mental health on social media platforms.

Model	F1-Score	Accuracy
Logistic Regression TFIDF	0.98	0.99
Logistic Regression W2Vec	0.63	0.87
Naïve Bayes TFIDF	0.9	0.96
DistilBERT	0.99	0.999

Figure 48: Comparative analysis of models based on F1-Score and Accuracy for Depression Twitter Dataset

Suicide Twitter Dataset

Logistic Regression TFIDF

Results for Logistic Regression with TF-IDF on Suicide Twitter Dataset: The classification report for the Logistic Regression model using TF-IDF vectorization on the Suicide Twitter Dataset shows high performance with an accuracy of 0.95. The precision, recall, and F1-score for class 0.0 (non-depressed) are 0.92, 1.00, and 0.96, respectively. For class 1.0 (depressed), the precision is 1.00, recall is 0.87, and the F1-score is 0.93. The macro average for precision, recall, and F1-score are 0.96, 0.93, and 0.94, respectively, while the weighted average for all metrics is 0.95.

Class	precision	recall	f1-score	support
0	0.92	1.0	0.96	110
1	1.0	0.87	0.93	68
accuracy			0.95	178
macro avg	0.96	0.93	0.94	178
weighted avg	0.95	0.95	0.95	178

Figure 49: Classification report for Logistic Regression TD-IDF on Suicide Twitter Dataset

The confusion matrix was also generated, highlighting the model's effectiveness. It shows 110 true positives and 59 true negatives, with 9 false negatives and no false positives. This indicates that the model is highly accurate in distinguishing between depressed and non-depressed tweets.

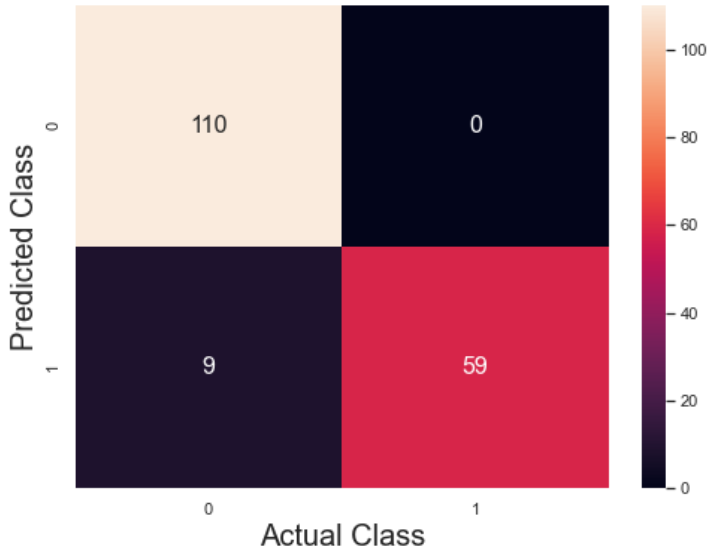

Figure 50: Confusion matrix for Logistic Regression TD-IDF on Suicide Twitter Dataset

The feature importance graph was produced, displaying the top 10 positive and negative features indicative of suicide and non-suicide. Terms like "die," "hate," "fack," and "damn" are strong indicators of suicide ideation, while words such as "thank," "happy," "lol," and "congratulation" are indicative of non-suicide ideation. This visualization provides insights into the key features that the model uses to make its predictions. This is a good indicator that the model performs with a good understanding of the task.

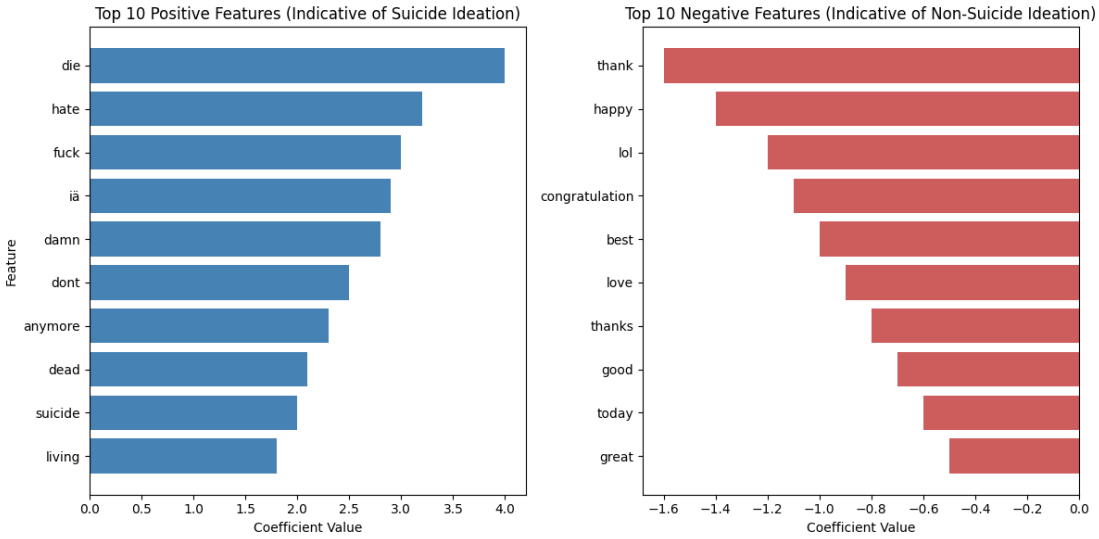

Figure 51: Top 10 positive and negative features for Logistic Regression TD-IDF on the Suicide Twitter Dataset

Naïve Bayes TFIDF

The classification report for the Naive Bayes model using TF-IDF vectorization on the Suicide Twitter Dataset shows high performance with an accuracy of 94%. The precision, recall, and F1-score for class 0 are 0.94, 0.96, and 0.95, respectively. For class 1 (suicide ideation), the precision is 0.94, recall is 0.90, and the F1-score is 0.92. The macro average for precision, recall, and F1-score are 0.94, 0.93, and 0.93, respectively, while the weighted average for all metrics is 0.94.

Class	precision	recall	f1-score	support
0	0.94	0.96	0.95	110
1	0.94	0.9	0.92	68
accuracy			0.94	178
macro avg	0.94	0.93	0.93	178
weighted avg	0.94	0.94	0.94	178

Figure 52: Classification report for Naïve Bayes TD-IDF on Suicide Twitter Dataset

The confusion matrix was also generated, highlighting the model's effectiveness. It shows 106 true positives and 61 true negatives, with 7 false negatives and 4 false positives. This indicates that the model is highly accurate in distinguishing between suicidal and non-suicidal tweets.

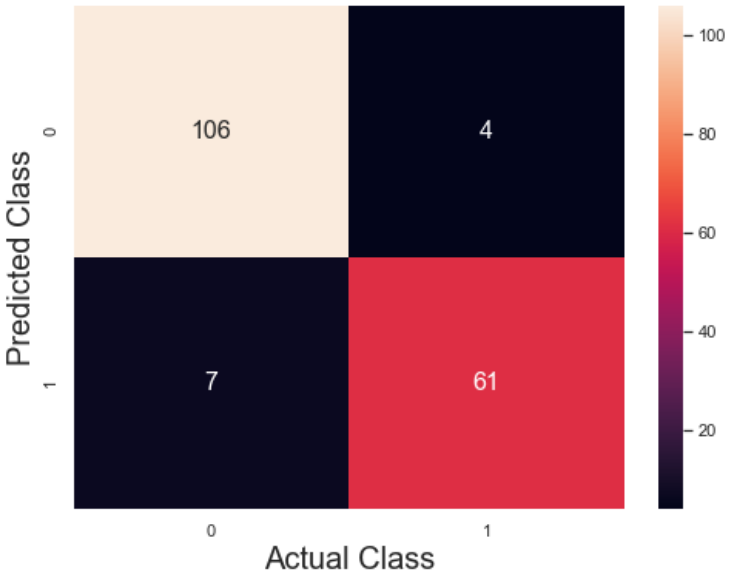

Figure 53: Confusion matrix for Naïve Bayes TD-IDF on Suicide Twitter Dataset

The feature importance graph was produced, displaying the top 10 positive and negative features indicative of suicide and non-suicide. Terms like "die," "hate," "fack," and "want" are strong indicators of suicide ideation, while words such as "bday," "aww," "award," and "appreciate" are indicative of non- suicide ideation.

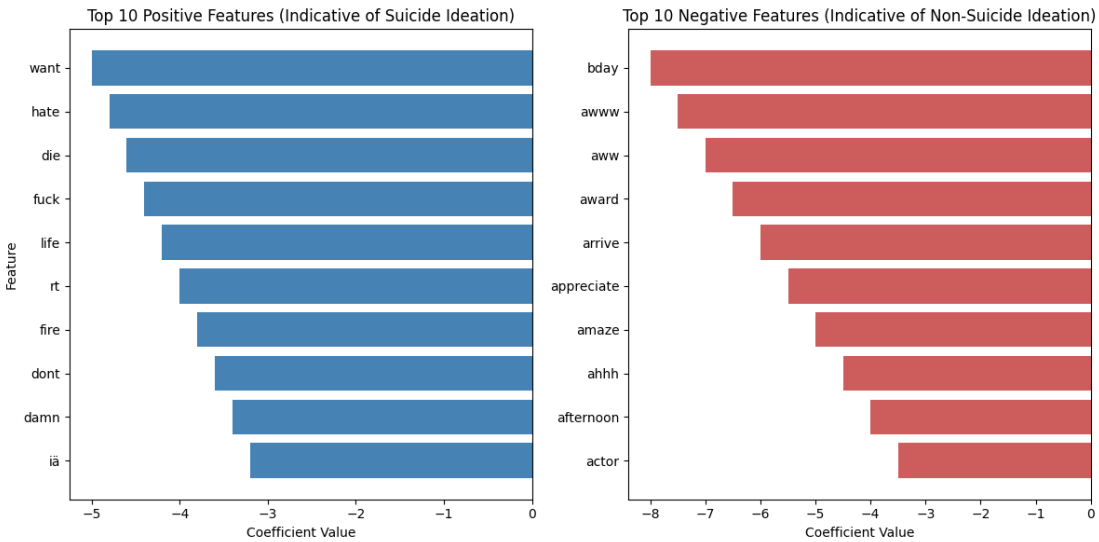

Figure 54: Top 10 positive and negative features for Naïve Bayes TD-IDF on the Suicide Twitter Dataset

Logistic Regression W2Vec

The classification report for the Logistic Regression model using Word2Vec embeddings on the Depression Twitter Dataset shows an overall accuracy of 61%. The precision, recall, and F1-score for class 0 are 0.62, 0.95, and 0.75, respectively. For class 1, the

precision is 0.38, recall is 0.04, and the F1-score is 0.08. The macro average for precision, recall, and F1-score are 0.50, 0.50, and 0.41, respectively, while the weighted average 0.49- 0.61.

Class	precision	recall	f1-score	support
0	0.62	0.95	0.75	110
1	0.38	0.04	0.08	68
accuracy			0.61	178
macro avg	0.5	0.5	0.41	178
weighted avg	0.52	0.61	0.49	178

Figure 55: Classification report for Logistic Regression W2Vec on Suicide Twitter Dataset

The confusion matrix highlights the model's performance, showing 105 true positives and 3 true negatives. There are 65 false negatives and 5 false positives, indicating a really lower accuracy in distinguishing between suicidal and non-suicidal tweets compared to other models.

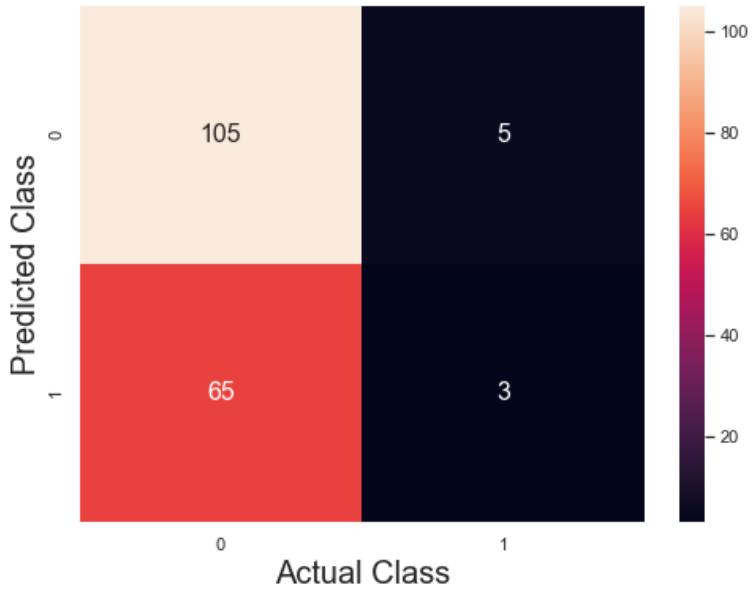

Figure 56: Confusion Matrix for Logistic Regression W2Vec on Suicide Twitter Dataset

Suicide BERT

This section summarizes the performance of the BERT model specifically trained to identify potential suicide-related posts. The BERT model achieved a precision of 0.89 for class 0 (non-suicide) and 0.86 for class 1 (suicide), with a recall of 0.86 and 0.90, respectively. The F1-score was consistently strong across both classes at 0.88. The model showed an overall accuracy of 0.88, with the macro average and weighted average for precision, recall, and F1-score all aligning at 0.88. This performance indicates the model's effectiveness in classifying suicide-related content accurately, maintaining a good balance between precision and recall.

	precision	recall	f1-score	support
0	0.89	0.86	0.88	226
1	0.86	0.9	0.88	225
accuracy			0.88	451
macro avg	0.88	0.88	0.88	451
weighted avg	0.88	0.88	0.88	451

Figure 57: Classification report for BERT model on Suicide Twitter Dataset

The confusion matrix provides a detailed look at the model's performance in classifying the posts. It shows that the model correctly identified 194 non-suicide posts and 202 suicide posts, while incorrectly classifying 32 non-suicide posts as suicide and 23 suicide posts as non-suicide. This performance underlines the model's higher sensitivity in identifying suicide posts with fewer false negatives, which is crucial for applications in preventive mental health strategies.

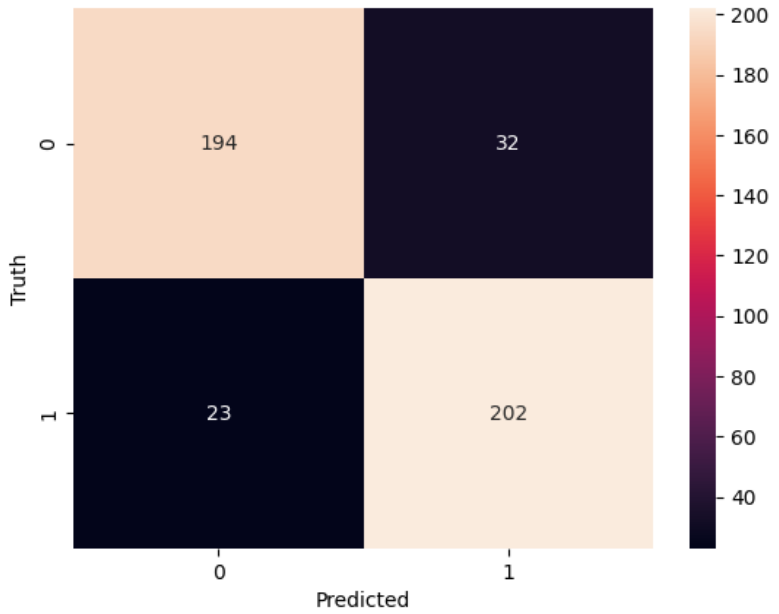

Figure 58: Classification report for BERT model on Suicide Twitter Dataset

Comparative Analysis Suicide Twitter Dataset

The addition of the BERT model to the comparison shows it achieving an accuracy and F1-score of 0.88. While not reaching the highest scores compared to Logistic Regression and Naïve Bayes, BERT's performance is still competitive, particularly given its capabilities in understanding contextual nuances in text. This suggests that while traditional models may excel in certain metrics, advanced models like BERT are valuable for their deep learning advantages, particularly in complex natural language processing tasks.

Model	F1-Score	Accuracy
Logistic Regression TFIDF	0.95	0.94
Logistic Regression W2Vec	0.41	0.63
Naïve Bayes TFIDF	0.93	0.94
BERT	0.88	0.88

Figure 59: Comparative analysis results table in Suicide Twitter Dataset

Chapter 5: Conclusions and Future Work

In this final chapter, we conclude our comprehensive analysis of various machine learning models applied to detect depression and suicidal content within social media posts. The research encompassed extensive data collection from platforms like Reddit and Twitter, followed by rigorous preprocessing, feature extraction, and model evaluation phases. We explored a range of traditional and advanced machine learning algorithms to assess their efficacy in recognizing mental health issues from textual data. Here, we summarize the key findings, discuss the implications of our results, and outline potential directions for future research in the burgeoning field of digital mental health monitoring.

Reddit Depression Dataset

The comparative analysis of different models used for classifying depression-related posts on the Reddit dataset reveals distinct levels of performance across the models. The Roberta model emerges as the top performer with an F1-score and accuracy of 0.98, showcasing its exceptional capability in handling natural language data with nuanced contexts. The Logistic Regression TFIDF model also demonstrates strong performance with an F1-score of 0.95 and an accuracy of 95%, indicating an effective balance between precision and recall. In contrast, the Naïve Bayes TFIDF model, with an F1-score of 0.89 and an accuracy of 88%, shows good effectiveness but is slightly outperformed by Roberta and Logistic Regression TFIDF. Lastly, the Logistic Regression W2Vec model displays relatively lower performance with an F1-score of 0.82 and an accuracy of 82%. This analysis underscores the varying effectiveness of different models in dealing with the intricacies of mental health-related text classification, highlighting the superior performance of models like Roberta and Logistic regression TFIDF in this domain.

Another conclusion comes from the feature importance graph produced, which illustrates the top 10 positive and negative features indicative of depression and non-depression. The top ten positive features strongly associated with depression include terms such as "depression," "anxiety," "feel," "life," "like," "get," "want," "go," "know," and "anymore." These terms are identified using methods like the Random Forest and Logistic Regression TFIDF, reflecting their high relevance and strong association with depressive posts

as indicated by their positive coefficient values. Conversely, the top ten negative features indicative of non-depression include lighter or neutral terms such as "haha," "window," "bar," "omg," "holiday," "ugh," "www," "homework," "update," and "rain." These terms, identified with significant negative coefficients, suggest their frequent occurrence in contexts that are typically not associated with depression. This refined analysis underscores the model's advanced ability to accurately understand and classify text based on the presence of 'positive' and 'negative' indicators relevant to mental health. The term "depression" consistently shows the highest positive coefficient, reaffirming its strong predictive value for depressive content. In contrast, terms like "haha" and "window" are highlighted as significant predictors of non-depressive content.

Furthermore, the analysis also underscores significant differences in message lengths between posts categorized as related to depression versus those that are not. Depression-related posts notably have a higher average length, exceeding 600 characters, suggesting that individuals discussing depression tend to write longer, more detailed posts to articulate their feelings, experiences, and struggles. In contrast, non-depression-related posts average around 100 characters, indicating a more concise form of communication. This discrepancy further illuminates the different communication patterns and needs between the two groups, emphasizing the depth of engagement and expression in depression-related discussions.

Twitter Depression Dataset

The comparative analysis of various models applied to the Depression Twitter Dataset reveals distinct performance levels in terms of F1-scores and accuracy rates. DistilBERT stands out with the highest F1-score of 0.99 and an impressive accuracy of 99.9%, indicating its superior performance in accurately identifying both depressed and non-depressed tweets. The Logistic Regression TFIDF model also demonstrates strong performance with an F1-score of 0.98 and an accuracy of 99%, reflecting an effective balance between precision and recall. In contrast, the Naïve Bayes TFIDF model, with an F1-score of 0.90 and an accuracy of 96%, shows good effectiveness but is slightly outperformed by DistilBERT and Logistic Regression TFIDF. Lastly, the Logistic Regression W2Vec model displays relatively lower performance with an F1-score of 0.63 and an accuracy of 87%. These findings highlight the high capability of advanced models like DistilBERT and Logistic Regression TFIDF in handling complex text classification tasks related to mental health on social media platforms, along with providing deeper insights into the linguistic and thematic elements of the tweets.

Moreover, the feature importance graph for this dataset shows the top 10 positive and negative features indicative of depression and non-depression. Terms like "depression," "anxiety," "pic," and "emoji" are strong indicators of depression, while words such as "movie," "tomorrow," "love," "thank," "bday," "bbq," "bb," and "bar" are indicative of non-depression. Again it underscores the model's nuanced understanding of language cues related to mental health.

In addition, the Hashtag Analysis plot reveals key insights into the discourse on Twitter. Frequently used hashtags such as #depression, #anxiety, #mentalhealth, #therapy, #stress, and #mentalillness indicate the primary topics of concern and related mental health issues. These hashtags are crucial for identifying specific mental health discussions, aiding in tracking mental health awareness campaigns, and understanding their impact on social media.

Lastly, analysis of tweet lengths and structures based on sentiment labels reveals again that tweets labeled as expressing depression tend to be significantly longer than non-depression tweets. This pattern suggests that individuals discussing depression use more words to articulate their feelings and experiences, potentially reflecting a greater need for expression and community support within social media platforms.

Twitter Suicide Dataset

BERT model achieving an accuracy and F1-score of 0.88. While not reaching the highest scores compared to models like Logistic Regression with TFIDF or Naïve Bayes with TFIDF, BERT's performance is still competitive, particularly given its capabilities in understanding contextual nuances in text. This suggests that while traditional models may excel in certain metrics, advanced models like BERT are valuable for their deep learning advantages, particularly in complex natural language processing tasks.

The models tested for classification included Logistic Regression with TFIDF, Logistic Regression with Word2Vec, Naïve Bayes with TFIDF, and BERT. Logistic Regression with TFIDF achieved the highest accuracy and F1-score, demonstrating its robustness in feature handling and decision-making capabilities. Naïve Bayes with TFIDF also performed well, showing significant strengths in handling text data. Logistic Regression with Word2Vec showed lower performance in comparison.

Despite scoring lower than some traditional models, BERT showed significant strengths in handling complex language nuances, crucial for real-world applications where context and semantic understanding are vital. This highlights the potential of pre-trained language models for nuanced semantic analysis, especially in identifying subtle indications in text data.

Moreover The analysis of the Twitter dataset for suicide ideation reveals significant insights into language use and message length in tweets labeled as 'Potential Suicide posts' and 'Not Suicide posts'. For tweets flagged as potential suicide, common words often express distress and negative emotions such as 'want,' 'don't,' 'hate,' 'life,' and 'tired,' which may indicate suicidal ideation. Notably, non-contextual words like 'https' and 'co' appear frequently due to their presence in URLs, necessitating removal during preprocessing to focus on meaningful content. In contrast, tweets labeled as 'Not Suicide post' predominantly feature positive and neutral terms like 'good,' 'want,' 'like,' 'love,' and 'happy.' These words reflect more typical everyday emotions and activities, underscoring a distinctly different tone and content compared to potential suicide posts.

Further analysis of bigrams—pairs of consecutive words—highlights phrases like 'want to,' 'to be,' 'tired of,' and 'my life' as common in the dataset, with some particularly distressing phrases such as 'hate myself,' 'kill myself,' and 'to die' underscoring severe emotional states and potential suicidal intentions. This indicates prevalent themes of existential reflection, fatigue, and severe distress in discussions around suicide.

Additionally, message length analysis reveals that tweets related to potential suicide intentions are significantly longer, averaging around 160 characters, compared to about 80 characters for non-suicide posts. This pattern is consistent with previous observations in depression-related analyses, where longer messages are used to articulate complex feelings and seek support, highlighting both the complexity and urgency of the communicators' emotional state.

These findings underscore the distinct linguistic patterns and emotional expressions in tweets associated with suicide ideation, providing valuable insights for potential preventive measures and support mechanisms on social media platforms.

General

This research aimed to develop effective machine learning models for identifying depression and suicidal content in social media posts, focusing on datasets from Twitter and Reddit. The methodology encompassed comprehensive data preprocessing, feature extraction, model training, and performance evaluation. A variety of models were explored, including traditional machine learning algorithms and advanced neural networks like

DistilBERT and ROBERTA, to ascertain their efficacy in classifying depressive and suicidal sentiments.

The study underscored the pivotal role of data preprocessing and feature extraction in amplifying model performance. Techniques such as tokenization, lemmatization, and TF-IDF vectorization were instrumental in enhancing the models' capacity to isolate relevant features from the noisy backdrop of social media data. Notably, the TF-IDF NLP method outperformed the W2VEC method, demonstrating a higher efficacy in handling the nuances of textual data in mental health contexts.

A common observation across the datasets was that posts from users exhibiting depressive symptoms or suicide ideation tended to be lengthier, possibly as an attempt to obfuscate their emotions or delve deeper into their experiences. This pattern underscores the complexity of mental health discussions on social media platforms and highlights the need for sophisticated analytical tools to interpret such expressions accurately.

Both Twitter and Reddit have proven to be invaluable resources for mental health monitoring. These platforms offer a real-time view into the public and private expressions of mental health, providing a rich dataset for training models that can detect nuanced emotional states. The findings suggest that while traditional machine learning models retain their relevance for certain applications, the integration of advanced models like DistilBERT and ROBERTA could significantly propel sentiment analysis forward, especially in the intricate domain of mental health assessment.

Finally, a key conclusion can be provided by the revised analysis of the common positive and negative words from the given datasets revealing distinct patterns. As we can observe in Figure 60, positive words indicative of depression or suicide ideation include "depression" and "anxiety." "Depression" is the most frequently associated word, appearing in all datasets, reflecting its centrality in the context, while "anxiety" frequently co-occurs with depression, indicating a common emotional state in these discussions.

Negative words indicative of non-depression or non-suicide ideation include "haha," "OMG," "love," and "thanks." "Haha" indicates laughter or amusement, commonly used in non-depressive contexts. "OMG" is an expression of surprise or excitement, not typically associated with depressive states. "Love" represents a positive emotional state, generally opposite to depression. "Thanks" is an expression of gratitude, indicative of positive interaction.

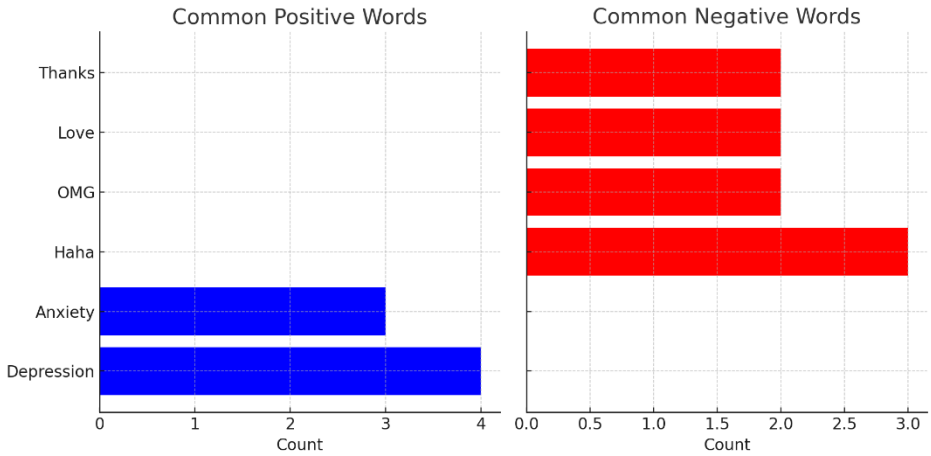

Figure 60: Common positve and negative words used between datasets that relates to depression or not-depression seperately

In addition to these terms, slang words such as "bday" (birthday), "bb" (baby or best buddy), "aww" (expression of endearment or sympathy), and "lol" (laughing out loud) are also indicative of non-depressive or non-suicidal states. These words are often used in casual, affectionate, or humorous contexts and reflect positive or neutral interactions.

It is interesting to note that despite analyzing different datasets, the models detect similar words as indicators. This consistency across varied data sources reinforces the reliability of these words as markers for depression, anxiety, and non-depressive states. The findings suggest that words indicative of depression and suicidal ideation are heavily weighted towards intense emotional expressions and states of mind, such as "depression" and "anxiety." In contrast, words indicative of non-depression or non-suicidal ideation include slang and casual expressions like "haha," "OMG," "bday," "bb," "aww," and "lol," as well as positive emotional states such as "love" and "thanks." These patterns reflect the stark contrast between discussions around depressive or suicidal states and more neutral or positive conversations.

This research contributes to the broader understanding of how machine learning can be harnessed to monitor mental health on social media effectively, providing insights that could facilitate early intervention and support for individuals displaying signs of depression or suicidal ideation.

Future research should explore integrating multimodal data sources, such as combining textual data with metadata or visual cues, to enrich the models' understanding and accuracy. Continuous adaptation and fine-tuning of models to new data and emerging linguistic trends will also be essential to maintain and improve performance over time. Additionally, developing more sophisticated techniques to handle imbalanced datasets, which are common in mental health-related data, could further enhance the effectiveness of these models.

From the data, it is evident that advanced models like BERT and ROBERTA, while not always outperforming traditional models in every metric, offer significant advantages in understanding and analyzing complex and nuanced text. The RandomForestClassifier's superior performance suggests that ensemble methods are particularly effective for classification tasks involving diverse and imbalanced data, such as social media posts. The slightly lower performance of BERT in some tests highlights a potential area for improvement in fine-tuning these models specifically for mental health-related content. It also suggests the need for more specialized training datasets that better capture the language used in suicidal and depressive posts.

The comparison across different datasets (Twitter and Reddit) reveals that model performance can vary significantly depending on the platform. This suggests that models need to be tailored and possibly fine-tuned for each specific platform to achieve the best results.

In summary, this research demonstrates the substantial potential of machine learning and natural language processing in mental health monitoring and intervention. By leveraging both traditional and advanced models, we can better understand and classify mental health-related content on social media, contributing to more effective public health initiatives. The integration of these technologies into mental health services could provide timely and accurate insights, ultimately helping to identify individuals in need and offering support where it is most urgently required.

Based on the results of our research, we conclude that advanced models like BERT and ROBERTA, along with traditional ensemble methods such as RandomForestClassifier and BaggingClassifier, are effective in identifying depression and suicidal content on social media. However, there are several areas for further improvement and exploration to enhance the robustness, applicability, and ethical use of these models.

Integration of Multimodal Data: Our results highlight the effectiveness of text-based models, but future research should explore the integration of multimodal data sources, such as combining textual data with metadata, visual cues, and audio signals. This approach can provide a more comprehensive understanding of the context and nuances in social media

posts, potentially enhancing the accuracy and reliability of detecting depression and suicidal content.

Continuous Model Adaptation: The evolving nature of language and trends on social media suggests that continuous adaptation and fine-tuning of models to new data are essential. Regular updates to training datasets and periodic re-training of models will help capture emerging linguistic trends and changes in user behavior, ensuring that the models remain relevant and effective over time. Our results showed varying effectiveness of different models across datasets, indicating the need for ongoing model refinement.

Handling Imbalanced Datasets: Our research revealed that class imbalances, where instances of depressive or suicidal content are relatively rare, can impact model performance. Developing more sophisticated techniques such as advanced sampling methods, cost-sensitive learning, and synthetic data generation can further enhance model effectiveness by improving sensitivity to rare but critical instances of mental health crises.

Advanced Preprocessing Techniques: While current preprocessing techniques such as tokenization, lemmatization, and TF-IDF vectorization proved effective, there is room for improvement. Future work could investigate more advanced preprocessing methods, including contextual embeddings and domain-specific lexicons, to better capture the semantic and syntactic nuances of mental health-related language. Our findings suggest that enhanced preprocessing can significantly improve model performance.

Development of Specialized Training Datasets: Creating specialized training datasets that accurately reflect the language used in suicidal and depressive posts is vital. These datasets should encompass a diverse range of expressions, slang, and cultural references common in social media. Collaboration with mental health professionals and linguists can aid in developing annotated datasets that capture the complexities of mental health discourse. Our research indicates that better-tailored datasets can improve model accuracy and relevance.

Real-Time Detection and Intervention: Implementing real-time detection systems for mental health crises on social media platforms can provide timely alerts to mental health professionals, enabling rapid intervention and support for individuals at risk. Ensuring ethical use, including privacy considerations and user consent, is paramount in deploying such systems. The high performance of models in our research underscores the feasibility of real-time applications.

Multilingual and Cross-Cultural Analysis: Expanding models to support multiple languages and cross-cultural analysis can broaden their applicability and impact. Building models that accurately detect depressive and suicidal content across different linguistic and cultural contexts will make the technology more inclusive and effective globally. Our results from English datasets suggest potential for extending these models to other languages and cultural contexts.

Collaboration with Mental Health Professionals: Engaging with mental health professionals is crucial to ensure that models are both accurate and ethically sound. Feedback from clinicians and therapists can refine model predictions and guide responsible use. Collaborative efforts can also aid in developing intervention strategies based on model outputs. Our findings highlight the importance of expert input in model development and application.

User Feedback and Model Interpretability: Incorporating user feedback mechanisms can improve model performance and trustworthiness. Understanding user experiences and perceptions of model accuracy provides valuable insights for further refinement. Additionally, enhancing model interpretability allows users and mental health professionals to understand the reasoning behind predictions, fostering greater confidence in the technology. Our research suggests that user feedback and transparency are essential for successful implementation.

To effectively support at-risk individuals, the development of real-time detection systems is crucial. Chatbots, in particular, could be a viable solution for providing immediate alerts and support. These systems must be trustworthy and inclusive, which requires expanding their capabilities to accommodate multiple languages and cultural contexts. Additionally, collaboration with mental health professionals is vital to ensure the accuracy and relevance of the models. Incorporating user feedback is also crucial for refining these models continually.

While large language models (LLMs) have demonstrated superior performance in various applications, they tend to be less controlled compared to traditional NLP-developed applications. To establish trustworthiness in mental health applications, it is necessary to address these issues. This includes enhancing all the mentioned improvements and integrating multimodal data sources like textual content, metadata, visual, and audio cues. Such integrations and enhancements will help in creating more robust and reliable systems for detecting and responding to mental health needs in real-time.

In conclusion, while significant progress has been made in using machine learning and natural language processing to detect depression and suicidal content on social media, numerous opportunities for future work remain. Addressing these areas will further improve the accuracy, reliability, and ethical application of these technologies, ultimately contributing to better mental health monitoring and intervention strategies.

References

[1] N. A. J. @ Abd Rahim and K. S. Ku Johari, "The Impact of Social Media on Mental Health among Adolescents," *International Journal of Academic Research in Business and Social Sciences*, vol. 13, no. 12. 2023. doi: 10.6007/ijarbss/v13-i12/20297.

[2] G. Zalsman *et al.*, "Suicide prevention strategies revisited: 10-year systematic review. The Lancet Psychiatry. 2016 Jul;3(7):646–59," *Pmid:*, vol. 27289303.

[3] V. Tejaswini, K. S. Babu, and B. Sahoo, "Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 1, 2024, doi: 10.1145/3569580.

[4] M. K. Kabir, M. Islam, A. N. Binte Kabir, A. Haque, and K. Rhaman, "Detection of Depression Severity Using Bengali Social Media Posts on Mental Health: Study Using Natural Language Processing Techniques," *JMIR Form. Res.*, vol. 6, no. 9, pp. 1–13, 2022, doi: 10.2196/36118.

[5] R. K. Bhoge, S. A. Nagare, S. P. Mahajan, and P. S. Kor, "Depression Detection by Analyzing Social Media Post of User," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 4, pp. 2720–2724, 2022, doi: 10.22214/ijraset.2022.41874.

[6] N. B. V. Riblet, B. Shiner, Y. Young-Xu, and B. V. Watts, "Strategies to prevent death by suicide: Meta-analysis of randomised controlled trials," *Br. J. Psychiatry*, vol. 210, no. 6, pp. 396–402, 2017, doi: 10.1192/bjp.bp.116.187799.

[7] C. M. van der Feltz-Cornelis *et al.*, "Best practice elements of multilevel suicide prevention strategies: A review of systematic reviews," *Crisis*, vol. 32, no. 6, pp. 319–333, 2011, doi: 10.1027/0227-5910/a000109.

[8] J. M. Eagles, D. P. Carson, A. Begg, and S. A. Naji, "Suicide prevention: A study of patients' views," *Br. J. Psychiatry*, vol. 182, no. MAR., pp. 261–265, 2003, doi: 10.1192/bjp.182.3.261.

[9] D. Ostic *et al.*, "Effects of Social Media Use on Psychological Well-Being: A Mediated Model," *Front. Psychol.*, vol. 12, no. June, 2021, doi: 10.3389/fpsyg.2021.678766.

[10] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *npj Digit. Med.*, vol. 3, no. 1, 2020, doi: 10.1038/s41746-020-0233-7.

[11] B. Keles, N. McCrae, and A. Grealish, "A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents," *Int. J. Adolesc. Youth*, vol. 25, no. 1, pp. 79–93, 2020, doi:

10.1080/02673843.2019.1590851.
[12] S. Brooks, “Does personal social media usage affect efficiency and well-being?,” *Comput. Human Behav.*, vol. 46, pp. 26–37, 2015, doi: 10.1016/j.chb.2014.12.053.
[13] A. Ahmed *et al.*, “Machine learning models to detect anxiety and depression through social media: A scoping review,” *Comput. Methods Programs Biomed. Updat.*, vol. 2, no. September, p. 100066, 2022, doi: 10.1016/j.cmpbup.2022.100066.
[14] R. F. Santalia, W. Gunadi, and A. Setiadi, “The Effect Of Emotional Support and Informational Support On The Need For Relatedness and User’s Satisfaction With The Use Of Social Media,” pp. 4365–4372, 2023, doi: 10.46254/an12.20220837.
[15] J. C. Eichstaedt, “AI-Enabled depression prediction using social media,” no. February, pp. 1–6, 2021.
[16] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, “Natural language processing in mental health applications using non-clinical texts,” *Nat. Lang. Eng.*, vol. 23, no. 5, pp. 649–685, 2017, doi: 10.1017/S1351324916000383.
[17] N. H. Kim, J. M. Kim, D. M. Park, S. R. Ji, and J. W. Kim, “Analysis of depression in social media texts through the Patient Health Questionnaire-9 and natural language processing,” *Digit. Heal.*, vol. 8, 2022, doi: 10.1177/20552076221114204.
[18] Z. Lim Yam, Z. Hayati Abdullah, and N. Filzah Mohd Radzuan, “Depression Detection Based On Twitter Using NLP and Sentiment Analysis,” *Appl. Math. Comput. Intell.*, vol. 11, no. 1, 2022.
[19] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, “Natural language processing applied to mental illness detection: a narrative review,” *npj Digit. Med.*, vol. 5, no. 1, pp. 1–13, 2022, doi: 10.1038/s41746-022-00589-7.
[20] L. Gan, Y. Guo, and T. Yang, “Machine Learning for Depression Detection on Web and Social Media: A Systematic Review,” *Int. J. Semant. Web Inf. Syst.*, vol. 20, no. 1, pp. 1–28, 2024, doi: 10.4018/IJSWIS.342126.
[21] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, “Natural Language Processing of Social Media as Screening for Suicide Risk,” *Biomed. Inform. Insights*, vol. 10, p. 117822261879286, 2018, doi: 10.1177/1178222618792860.
[22] T. H. H. Aldhyani, S. N. Alsubari, A. S. Alshebami, H. Alkahtani, and Z. A. T. Ahmed, “Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, 2022, doi: 10.3390/ijerph191912635.
[23] V. Nasteski, “An overview of the supervised machine learning methods,” *Horizons.B*, vol. 4, no. December, pp. 51–62, 2017, doi: 10.20544/horizons.b.04.1.17.p05.
[24] J. Zou, T. Hirokawa, J. An, L. Huang, and J. Camm, “Recent advances in the applications of machine learning methods for heat exchanger modeling—a review,” *Front. Energy Res.*, vol. 11, no. November, pp. 1–25, 2023, doi: 10.3389/fenrg.2023.1294531.
[25] “Artemis_ ΑΥΤΟΜΑΤΗ ΑΝΑΓΝΩΡΙΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΜΕΣΩ ΣΥΝΔΥΑΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΦΥΣΙΟΛΟΓΙΚΩΝ ΚΑΙ ΣΥΜΠΕΡΙΦΟΡΙΚΩΝ ΠΑΡΑΜΕΤΡΩΝ.”
[26] X. Ying, “An Overview of Overfitting and its Solutions,” *J. Phys. Conf. Ser.*, vol. 1168, no. 2, 2019, doi: 10.1088/1742-6596/1168/2/022022.
[27] R. Teehan *et al.*, “Emergent Structures and Training Dynamics in Large Language Models Charles River Analytics 2 Edinburgh Centre for Robotics,” pp. 146–159, 2022.
[28] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, “Stress detection using natural language processing and machine learning over social interactions,” *J. Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00575-6.
[29] G. Neelavathi, D. Sowmiya, C. Sharmila, and J. Vaishnavi, “Sentiment Analysis for Depression Based on Social Media Post by Using Natural Language Processing,” *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 12, no. 2, pp. 134–139, 2021, doi:

10.48175/ijarsct-2319.
[30] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," 2023, [Online]. Available: http://arxiv.org/abs/2307.06435
[31] A. Radwan, M. Amarneh, H. Alawneh, H. I. Ashqar, A. AlSobeh, and A. A. A. R. Magableh, "Predictive Analytics in Mental Health Leveraging LLM Embeddings and Machine Learning Models for Social Media Analysis," *Int. J. Web Serv. Res.*, vol. 21, no. 1, pp. 1–22, 2024, doi: 10.4018/IJWSR.338222.
[32] G. E. Vaillant, "Mental health," *American Journal of Psychiatry*, vol. 160, no. 8. pp. 1373–1384, 2003. doi: 10.1176/appi.ajp.160.8.1373.
[33] A. Palanica, P. Flaschner, A. Thommandram, M. Li, and Y. Fossat, "Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey," *J. Med. Internet Res.*, vol. 21, no. 4, pp. 1–10, 2019, doi: 10.2196/12887.
[34] C. Sweeney *et al.*, "Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts," *ACM Trans. Comput. Healthc.*, vol. 2, no. 3, 2021, doi: 10.1145/3453175.
[35] "Scalable Enterprise Solution for Mental Health | Woebot Health." [Online]. Available: https://woebothealth.com/
[36] Wysa, "Wysa - Everyday Mental Health," *Wysa*. 2022. [Online]. Available: https://www.wysa.io/
[37] "Youper: Artificial Intelligence For Mental Health Care." 2024. [Online]. Available: https://www.youper.ai/
[38] "Marlee, the world's first AI Coach."
[39] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," *Proc. 14th Int. AAAI Conf. Web Soc. Media, ICWSM 2020*, no. January, pp. 830–839, 2020, doi: 10.1609/icwsm.v14i1.7347.
[40] D. N. Milne, G. Pink, B. Hachey, and R. A. Calvo, "CLPsych 2016 shared task: Triaging content in online peer-support forums," *Proc. 3rd Work. Comput. Linguist. Clin. Psychol. From Linguist. Signal to Clin. Reality, CLPsych 2016 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lan*, no. January, pp. 118–127, 2016, doi: 10.18653/v1/w16-0312.
[41] D. E. Losada and F. Crestani, "A Test Collection for Research on Depression and Language Use CLEF 2016, Évora (Portugal)," *Exp. IR Meets Multilinguality, Multimodality, Interact.*, pp. 28–29, 2016, [Online]. Available: https://tec.citius.usc.es/ir/pdf/evora.pdf
[42] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses," *2nd Comput. Linguist. Clin. Psychol. From Linguist. Signal to Clin. Reality, CLPsych 2015 - Proc. Work.*, no. December, pp. 1–10, 2015, doi: 10.3115/v1/w15-1201.
[43] X. Xu *et al.*, "Mental-LLM," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 1, pp. 1–32, 2024, doi: 10.1145/3643540.
[44] R. McKee, "Ethical issues in using social media for health and health care research," *Health Policy (New. York).*, vol. 110, no. 2–3, pp. 298–301, 2013, doi: 10.1016/j.healthpol.2013.02.006.
[45] R. F. Hunter *et al.*, "Ethical issues in social media research for public health," *American Journal of Public Health*, vol. 108, no. 3. pp. 343–348, 2018. doi: 10.2105/AJPH.2017.304249.
[46] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, 2022, doi: 10.1007/s11280-021-00992-2.

[47] M. Zimmer, "'But the data is already public': On the ethics of research in Facebook," *Ethics Inf. Technol.*, vol. 12, no. 4, pp. 313–325, 2010, doi: 10.1007/s10676-010-9227-5.

[48] A. Xin, J. A. Lossio-ventura, K. R. Krause, G. Fiorini, F. Pereira, and D. M. Nielson, "Outcomes that matter to depressed adolescents can be identified with large language models," pp. 1–26.

[49] R. Improta and M. Stella, "I NTRODUCING C OUNSE LLM E : A DATASET OF SIMULATED MENTAL HEALTH DIALOGUES FOR COMPARING LLM S LIKE H AIKU , LL A MA NTINO AND," pp. 1–19, 2024.

[50] Z. Elyoseph, I. Levkovich, and S. Shinan-Altman, "Assessing prognosis in depression: Comparing perspectives of AI models, mental health professionals and the general public," *Fam. Med. Community Heal.*, vol. 12, no. Suppl 1, 2024, doi: 10.1136/fmch-2023-002583.

[51] A. Pérez, M. Fernández-Pichel, J. Parapar, and D. E. Losada, "DepreSym: A Depression Symptom Annotated Corpus and the Role of LLMs as Assessors of Psychological Markers," pp. 1–5, 2023, [Online]. Available: http://arxiv.org/abs/2308.10758

[52] N. Farruque, R. Goebel, S. Sivapalan, and O. R. Zaïane, "Depression symptoms modeling from social media text: an LLM driven semi-supervised learning approach," *Lang. Resour. Eval.*, 2024, doi: 10.1007/s10579-024-09720-4.

[53] J. Ohse *et al.*, "Zero-shot strike: Testing the generalization capabilities of out-of-the-box LLM models for depression detection," *Comput. Speech Lang.*, vol. 88, no. April, p. 101663, 2024, doi: 10.1016/j.csl.2024.101663.

[54] N. Wang *et al.*, "Learning Models for Suicide Prediction from Social Media Posts," *Comput. Linguist. Clin. Psychol. Improv. Access, CLPsych 2021 - Proc. 7th Work. conjunction with NAACL 2021*, pp. 87–92, 2021, doi: 10.18653/v1/2021.clpsych-1.9.

[55] T. M. Fonseka, V. Bhat, and S. H. Kennedy, "The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors," *Aust. N. Z. J. Psychiatry*, vol. 53, no. 10, pp. 954–964, 2019, doi: 10.1177/0004867419864428.

[56] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," *Algorithms*, vol. 13, no. 1, pp. 1–19, 2020, doi: 10.3390/a13010007.

[57] A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z. A. Kaminsky, "A machine learning approach predicts future risk to suicidal ideation from social media data," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–12, 2020, doi: 10.1038/s41746-020-0287-6.

[58] T. B. Brown *et al.*, "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, 2020.

[59] C. A. Bejan *et al.*, "Improving ascertainment of suicidal ideation and suicide attempt with natural language processing," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-19358-3.

[60] K. Gupta, R. Jinad, and Q. Liu, "Comparative Analysis of NLP Models for Detecting Depression on Twitter," *Proc. - 2023 Int. Conf. Commun. Comput. Artif. Intell. CCCAI 2023*, pp. 23–28, 2023, doi: 10.1109/CCCAI59026.2023.00013.

[61] R. Rachh and S. Kavatagi, "Performance Evaluation of Various Embedding Techniques for Identification of Depressive Contents in Social Media," *2023 3rd Asian Conf. Innov. Technol. ASIANCON 2023*, pp. 1–6, 2023, doi: 10.1109/ASIANCON58793.2023.10270579.

[62] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T. S. Chung, "A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/7893775.

[63] "Depression: Reddit Dataset (Cleaned)." 2022. [Online]. Available: https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned

[64] G. Manas, “Sentimental Analysis for Tweets | Kaggle,” *Kaggle*. 2021.
[65] M. Z. Islam, S. A. Mahmud, and N. Islam, “Suicidal Tweet Detection Dataset,” *Kaggle*. 2023. [Online]. Available: https://www.kaggle.com/datasets/aunanya875/suicidal-tweet-detection-dataset/suggestions?status=pending&yourSuggestions=true
[66] C. Ashby and D. Weir, “Leveraging HTML in Free Text Web Named Entity Recognition,” *COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 407–413, 2020, doi: 10.18653/v1/2020.coling-main.36.
[67] Daniel Jurafsky & James H. Martin, “Hidden Markov and Maximum Entropy Dra,” *Speech Lang. Process. An Introd. to Nat. Lang. Process. Comput. Linguist. speech Recognit.*, p. 1004, 2006, [Online]. Available: papers3://publication/uuid/531A3835-7600-4448-853F-34C2CDA40E8D
[68] M. Das, S. Kamalanathan, and P. Alphonse, “A Comparative Study on TF-IDF feature weighting method and its analysis using unstructured dataset,” *CEUR Workshop Proc.*, vol. 2870, pp. 98–107, 2021.
[69] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
[70] “Depressed tweet detection DistilBERT.”
[71] “Suicidal Tweet Detection using Multiple ML Model.”
[72] “Suicide Tweet Detection Using Bert.”