ABSTRACT

The contemporary landscape of social media, epitomized by platforms like X (previously known as Twitter), serves as a dynamic arena where communication thrives alongside myriad challenges, particularly concerning mental well-being and user safety. This study conducts a comprehensive review of existing research, focusing on the detection and intervention methods for instances of self-harm and expressions of suicidal ideation within the realm of X.

In the digital age, social media platforms have become integral to communication, facilitating instantaneous connection and information dissemination. However, this interconnectedness also brings to the fore concerning issues related to mental health, including cyberbullying, self-harm, and suicidal thoughts. Recognizing the gravity of these challenges, the primary aim of this review is to critically evaluate the various sentiment analysis and language processing techniques employed in prior studies to identify and address such distress signals within the vast sea of social media data.

By delving into the existing methodologies, this review seeks not only to understand their efficacy but also to propose innovative approaches to augment current practices. The synthesis of findings from diverse studies enables a nuanced understanding of the strengths and limitations of existing procedures, paving the way for the development of more robust strategies for detection and intervention.

Crucially, the conclusions drawn from this review advocate for heightened research efforts aimed at implementing measures that mitigate internet vulnerabilities. By bolstering the security infrastructure of platforms like X and enhancing awareness about mental health issues, strides can be made towards fostering a safer and more supportive online environment.

Moreover, the analysis conducted in this study extends beyond technical considerations to explore potential partnerships with legal entities and online communities. These collaborations offer promising avenues for implementing holistic solutions that address not only the symptoms but also the underlying causes of online distress, such as cyberbullying and self-harm.

Drawing insights from survey data, it becomes apparent that ensuring user safety and comfort online necessitates a multifaceted approach. While technological interventions play a pivotal role, cultivating a culture of empathy, respect, and digital responsibility within online communities is equally essential.

In essence, this study underscores the imperative for ongoing research and collaborative action to safeguard the mental well-being of social media users. By harnessing the power of sentiment analysis and leveraging interdisciplinary partnerships, we can aspire towards a digital landscape where individuals feel supported, empowered, and resilient in the face of online challenges.

Chapter-1

INTRODUCTION

1.1 BACKGROUND

In the face of rapid change in the modern-day digital world social media platforms have taken on the mantle of communication across geographical differences. Nevertheless, the advent of the digital age carries a set of problems, especially mental health issues and security. Similarly to the detection and intervention of online dangers, the issue of self-harm and the manifestation of suicidal ideas has become an important concern in terms of e-safety.

The 21st century witnessed the web as having become the fourth industrial revolution. It corrected communication, trade, and even political discourse. Every day things create a lot of data that is hard to believe. We can expect the amount of data generated by every person on this planet to grow to 1.7 megabytes from every internet user by 2025. The yearly exponential growth of the human species becomes greater than the entire time of mankind.

This review paper will uncover the multitude of web threats that occur on the X platform (formerly known as Twitter). Twitter, best known for its microblogging component, is capable of transforming even the most elaborate language into a 280-character language or even fewer characters. And with a lot of users willing to share with others their feelings, emotions, and attitudes X gives a unique sample that helps to analyze the results.

Underlying attention is invested in sentiment analysis. We have an objective to simplify machine learning and natural language processing (NLP), and from this process, we will be able to get the emotional undercurrent of X. Are we able to distinguish joy, frustration, or despair from an anonymous message? How do users vent their moods and what types of moods and patterns appear?

Unlike finding new tools, our aim is more focused on a deep understanding of what we will find by using the existing methods. We go deep shift into lexicons with sentiment, machine learning algorithms, and semantically associated terms. Through the analysis of X, getting to know the emotional pulse of online communications helps us to weave a comprehensive online discourse.

It is more about X as such. It acts as a knowledge base for subsequent platforms to use replicating effective strategies and tactics to not only curb cyberbullying but also encourage users' mental well-

being. In the course of it, can we make ourselves more responsible and merciful—at a time when the digital space has become full of information.

1.2 PROBLEM STATEMENT

The objective of this review is to conduct a comprehensive assessment of two primary aspects: firstly, the existing methods utilized for the detection of suicidal and self-harming language within past iterations of Twitter, and secondly, the strategies employed to address such instances. Emphasizing the adoption of diverse approaches within the domains of mental health awareness and online safety, this review seeks to underscore their significant contribution towards enhancing overall awareness and safety measures concerning mental health and online well-being. By critically evaluating these aspects, this review aims to inform and advance efforts aimed at promoting a healthier and safer digital environment for all users.

1.3 AIMS AND OBJECTIVES

The primary objective of this project is to develop an innovative and impactful tool with the core purpose of identifying and addressing critical issues on Twitter, specifically cyber bullying, self-harm, and expressions of suicidal thoughts. The system is designed to provide timely support to individuals in distress, ultimately contributing to the broader mission of preventing self-harm and promoting mental health awareness within the online community.

This project aspires to create a pioneering solution that not only detects harmful content but also responds effectively, demonstrating a commitment to the well-being and safety of Twitter users. The system's capabilities extend to early intervention, crisis management, and resource provisioning, ensuring that individuals facing distressing situations receive the support they need in a timely manner.

Furthermore, the successful implementation of this system may set a precedent and offer an exemplary model for similar platforms and online communities to adopt. By demonstrating the effectiveness of this tool, the project aims to influence and encourage the integration of similar mechanisms across the digital landscape, thereby enhancing online safety and contributing to the collective mission of fostering mental health awareness.

1.4 THESIS LAYOUT

Basically, the thesis is consisted of **six sections**;

Section 1: This review paper delves into the complexities of detecting and addressing mental health issues, particularly self-harm and suicidal ideation, within the digital realm, focusing on Twitter (referred to as X). It aims to analyze existing methodologies and strategies while proposing an innovative tool to combat cyberbullying and support mental well-being. Ultimately, the project seeks to pioneer a solution that not only identifies harmful content but also provides timely support, potentially setting a precedent for similar platforms to prioritize online safety and mental health awareness. **Section 2**: Brief History of Work Done This chapter reviews existing literature on sentiment analysis and related topics within the context of Twitter. It highlights a diverse range of studies focusing on detecting self-harm tendencies, analyzing crime-related tweets, and understanding discussions about non-suicidal self-injury and suicidal ideation. Various methodologies, including machine learning algorithms, deep learning techniques, and sentiment analysis, are explored, showcasing advancements and challenges in leveraging Twitter data for mental health awareness and online safety. **Section 3:** This section presents a comprehensive review of 20 papers in the field, focusing on synthesizing their methodologies and performance metrics, including F-measure, precision,

and recall. By analyzing each paper's experimental setups and reported results, the review aims to provide insights into the advancements and challenges in sentiment analysis techniques. Leveraging the confusion matrix, performance metrics are derived to assess model accuracy, precision, and recall, offering a nuanced understanding of their comparative performance and suitability for different applications in the field. **Section 4:** In the results and analysis, various machine learning algorithms and methods were applied across different research papers to address sentiment analysis and depression identification. Techniques such as Naive Bayes, Random Forest, LSTM-CNN, and hybrid approaches combining word embeddings and deep learning achieved notable accuracies ranging from 59% to 100%. Precision, recall, and F-measure were also reported, indicating the effectiveness of different models in capturing relevant instances and minimizing false positives. Overall, the findings showcase the diversity of approaches and their respective performances in tackling sentiment analysis and depression detection tasks.**Section 5**: The discussion and conclusion of this review paper underscore the importance of sentiment analysis, particularly in social media platforms like Twitter, for various applications such as identifying self-harm tendencies and understanding crime-related discussions. The diverse methodologies employed, ranging from traditional machine learning to advanced deep learning models, showcase high accuracy rates, emphasizing their effectiveness in addressing complex challenges. These findings highlight the interdisciplinary significance of sentiment analysis in enhancing mental health awareness, public safety, and overall well-being in the digital landscape, providing a foundation for future research and proactive measures **Section 6**: The summary chapter outlines the objectives of the project and evaluates the degree of success achieved, citing relevant facts and figures. It highlights the importance of the research in advancing sentiment analysis techniques and improving understanding of user sentiments in social media data. Additionally, it suggests future avenues for research, such as interpreting sarcasm, integrating multimodal data, real-time analysis, cross-platform generalization, and human-AI collaboration, to further enhance the field's capabilities and applications.

Chapter-2

LITERATURE REVIEW

2.1 Brief History of Work Done

X, a microblogging platform with over 528 million monthly active users, has emerged as a valuable data source for sentiment analysis. This literature review provides an overview of existing research on Twitter sentiment analysis, highlighting key themes, methodologies, and recent trends in this field.

[1]Rinku Yadav, Varun Gupta*(2018).Self-Harm Prevention Based On Social Platforms User Data Using Naive Bayes Classifier textblob.

This research addresses sentiment analysis, specifically focusing on identifying self-harm tendencies using Twitter data. Existing approaches can be grouped into knowledge-based, statistical, and hybrid techniques. Knowledge-based methods use prior knowledge and logical operations to predict sentiment, while statistical methods involve feature extraction from text for prediction. Hybrid approaches combine both. Analyzing social platform text poses challenges like word order, grammatical errors, and user-specific abbreviations, which can provide insights into self-harm tendencies.

The research uses a dataset of 200+ entries from Twitter, including usernames, tweet text, and sentiment labels (positive, negative, neutral). It employs a naive Bayes classifier from the text blob library trained with positive and negative words. The goal is to detect self-harm tendencies based on sentiment analysis, allowing timely intervention and counseling to prevent self-harm incidents. The accuracy obtained during the training and testing phase is 100% and 82.2 % respectively.

[2]Sangeeta Lal, Lipika Tiwaria, Ravi Ranjana, Ayushi Verma, Neetu Sardanaa, Rahul Mourya(2020).Analysis and Classification of Crime Tweets.

Scholarly investigations have begun to recognize the potential of this user-generated crime-related data on Twitter. Such data can be harnessed for enhancing crime management and response strategies. By systematically analyzing and classifying crime-related tweets, it becomes possible to identify genuine calls for police intervention. This capability aligns with the broader trend of utilizing social media data for public safety and law enforcement.

The pivotal role of classification tools is underscored in this context. These tools are essential for efficiently categorizing crime-related tweets, enabling police departments to optimize their resource allocation and focus on tweets that genuinely require intervention.

The study at hand builds upon this evolving body of knowledge. The research employs a text

mining-based approach to classify a dataset of 369 tweets into two categories: those related to crimes requiring police attention and those that are not crime-related.

The analysis involves a comparison of four classifiers: Naive Bayesian (NB), Random Forest, J48, and ZeroR. The results reveal that the Random Forest (RF) classifier outperforms the others, achieving an accuracy rate of 98.1%. In contrast, the ZeroR classifier is the least effective, with an accuracy rate of 61.5%.

[3]Muhammad Abubakar Alhassan, and Diane Pennington(2021).Investigating Non-suicidal Self-injury Discussions on Twitter.

The study analyzed sentiments, examined conversation themes, and classified participants based on their participation in NSSI debates on Twitter. It acknowledged that, in contrast to earlier visual content studies, a thorough analysis of textual NSSI content on Twitter was necessary.

The study used Twitter Archive Google Sheets (TAGS) to gather data by using the hashtags #selfharm and #selfinjury. Text processing, user classification, sentiment analysis, and LDA topic modeling were all done. The study determined that self-harm misjudgment, mental health awareness, school, and mental health assistance, suicide and mental health difficulties, and children and youth well-being were the five main subjects discussed in NSSI talks.

The results of the survey showed that different user groups expressed different sentiments. Discussions were positively impacted by academic professionals, medical teams, and support groups, whereas non-professional users expressed more unfavorable emotions. The main subjects of conversation were mental health awareness, school assistance, and busting myths about self-harm.

[4]E. Rajesh Kumar, N. Venkatram(2023).A novel approach for Communication related to suicidal detection on Twitter using multi-class data.

This report provides an overview of a research study that focuses on identifying early indicators of suicidal thoughts in Twitter posts through text data analysis. The research underscores the importance of analyzing textual content on social media platforms, like Twitter, to address the global public health problem of suicide. Previous studies have shown an increasing interest in using social media data to detect signs of suicidal behavior and mental health issues, leveraging Natural Language Processing (NLP) and text analysis techniques.

The study collected data from Twitter by mining posts with specific hashtags related to self-harm and self-injury over a year-long period, starting from February 1, 2018, resulting in a substantial dataset of over two million posts.

The Twitter data was categorized into seven different classes, identifying various types of suicide-related communication, including evidence of suicide attempts, support or information, flippant references, campaigns or fights, condolences or memorials, suicide reporting, and none of these. Three sets of features were extracted from the tweets, covering structural, lexical, psychological,

and emotional attributes, crucial for the subsequent classification process.

The research employed both baseline and ensemble classification methods. Baseline classifiers included a rule-based classifier, Naive Bayes classifier, and Support Vector Machine (SVM). Additionally, an ensemble model using Rotation Forest (RF) was introduced to enhance classification accuracy through feature diversity.

The study reported an F-measure of 0.82 for all seven classes, with a specific F-measure of 0.76 for suicidal ideation, signifying a significant improvement in classifying tweets containing suicidal thoughts. The research identified essential linguistic patterns and features associated with different types of suicide-related communication.

This study represents a significant step towards identifying and understanding suicidal language on Twitter, with the potential to contribute to preventive strategies and support for individuals struggling with suicidal thoughts. It underscores the value of social media data analysis in addressing mental health concerns, particularly related to suicidal ideation. The research concluded by emphasizing the importance of collaboration with experts in suicidology and suggested the possibility of extending the method to other social media platforms and text sources.

[5] Selva Mary G., John Blesswin A., Mithra Venkatesan, Shubhangi Vairagar, Sushadevi Adagale, Chetana Shravage, Jyotsna Barpute(2023).Enhancing conversational sentimental analysis for psychological depression prediction with Bi-LSTM.

This study focuses on the early detection of depression by analyzing user-generated content on major social media platforms like Twitter, Facebook, and Instagram. The research employs natural language processing and machine learning, using sentiment analysis to interpret emotional context in posts and comments.

A novel methodology utilizing Bidirectional Encoder Representations from Transformers (BERT) is proposed for efficient analysis. Knowledge distillation enhances accuracy by transferring insights from a large BERT model to a smaller one. Integrating word2vec and BERT with bidirectional long short-term memory (Bi-LSTM), the approach effectively identifies depression and anxiety indicators in social media data.

The proposed BERT-Bi-LSTM model outperformed other machine learning classification models by effectively capturing the syntactic and contextual information of each word. The study's methodology offers significant potential for early detection of mental health issues and personalized mental health care strategies.

[6] Krishna Shrestha (2018). Machine Learning for Depression Diagnosis using Twitter data

The document reviews previous studies that have employed machine learning techniques to detect depression. It delves into the data collection methods used, the choice of depression scales for evaluation, and the feature extraction techniques applied. Various studies have adopted different strategies for data collection, including mining public tweets and conducting surveys. The choice of

depression scale, such as the Centre for Epidemiological Studies Depression (CES-D) Scale, plays a vital role in assessing the degree of depression in users. Depending upon the CES-D score, the likelihood of depression may be: low (0-15), mild to moderate (16-22) or high (23-60)Feature extraction methods vary across studies, with some using sentiment analysis, linguistic inquiry, and word frequency analysis to generate relevant features.

[7] Michael Mesfin Tadesse, Hongfei Lin, Bo Xu , and Liang Yang(2019). Detection of Suicide Ideation in Social Media Forums Using Deep Learning

The paper presents a study on suicide ideation detection in social media, particularly on Reddit. The authors use various natural language processing and text classification techniques, including a combination of LSTM and CNN models, to improve suicide ideation detection. They analyze data from suicide-indicative and non-suicidal posts, identifying linguistic patterns associated with suicidal thoughts. The proposed LSTM-CNN model outperforms other machine learning classifiers, achieving a high accuracy of 93.8% and an F1 score of 93.4%. The study discusses the importance of CNN in text classification and highlights the limitations related to data deficiency and annotation bias. The authors suggest that their research could contribute to building effective suicide detection and reporting systems on social media.

[8]Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long (2018).Supervised Learning for Suicidal Ideation Detection in Online User Content.

The paper compared different classification methods, including Support Vector Machine (SVM), Random Forest, Gradient Boosting, XGBoost, Multilayer Feed-Forward Neural Network (MLFFNN), and Long Short-Term Memory (LSTM) networks. These models were trained to distinguish between suicidal and non-suicidal posts based on the extracted features.

The authors conducted experiments on Reddit and Twitter datasets to evaluate the performance of the classification models. They used metrics such as accuracy, precision, recall, F1-score, and AUC (Area Under the Curve) to assess the models' effectiveness. The results showed that combining multiple features led to improved performance in identifying suicidal ideation, with XGBoost achieving the highest AUC.

The paper also compared the classification of suicidal posts with posts from specific subreddits, such as "gaming," "jokes," "books," "movies," and "AskReddit." Surprisingly, the models achieved better results in distinguishing suicidal posts from other subreddit topics than from general non-suicidal posts.

The authors extended their experiments to Twitter data, where they applied feature processing and classification models. They addressed class imbalance issues by using undersampling techniques and compared the performance of different models.

[9] Minsu Park, Chiyoung Cha, and Meeyoung Cha(2012). Depressive Moods of Users Portrayed in Twitter

This study investigates the relationship between depression and social media posts, specifically on Twitter. The research seeks to understand how individuals express and share their experiences related to depression on this platform. The study collected data from 69 young adults who underwent a depression screening process and were allowed access to their social media posts. Sentiment analysis was applied to the tweets to explore the connection between language usage and depression. The results revealed that 28 participants had a high probability of depression, while 41 had low or mild depression. Depressed users tended to express more negative emotions and anger in their tweets. Interestingly, the study did not find gender or age to be significant factors in predicting depression, although previous research has suggested otherwise.

The discussion section emphasizes the implications of using social media data for clinical studies on depression. It points out the potential for real-time healthcare support and the importance of understanding how individuals express their emotions online. The study acknowledges its limitations, such as the small sample size and the need to adapt sentiment analysis tools to evolving language. In conclusion, the study suggests that online social network data, particularly from platforms like Twitter, can offer valuable insights into the study of depression. It holds promise for improving our understanding of depression and developing real-time support systems for individuals dealing with this mental health condition.

[10] Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat and A. Rehman (2017).Sentiment Analysis Using Deep Learning Techniques: A Review

The studies you provided highlight the effectiveness of deep learning models in the field of sentiment analysis. Deep learning models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Deep Belief Networks (DBNs), and others, have demonstrated superior performance in sentiment analysis tasks. They offer advantages such as automatic feature extraction, the ability to handle various problem statements, and the potential for both supervised and unsupervised learning.

These models have outperformed traditional methods like Support Vector Machines (SVMs) and shallow neural networks, providing higher accuracy in sentiment classification. They are well-suited for tasks that involve large datasets and can handle complex variations in the sentiment analysis process.

However, it's important to note that deep learning models come with certain limitations, including the need for substantial computational resources and extensive training times, especially when using GPUs. Despite these challenges, their ability to accurately analyze and predict sentiment makes them valuable tools in the field of sentiment analysis.

[11] Pete Burnapa , Gualtiero Colomboa, Rosie Amery b , Andrei Hodoroga , Jonathan

Scourfieldc (2017). Multi-class machine classification of suicide-related communication on Twitter

The main objective of this research is to create machine learning models that can classify textual material on suicide into several groups, such as suicidal thoughts, suicide reporting, memorial posts, campaigning, support, and finding casual allusions to suicide. Using a range of data taken from Twitter messages, including lexical, structural, emotive, and psychological characteristics, the authors begin by building baseline classifiers. The performance of these basis classifiers is then improved by building an ensemble classifier that combines the output of the base classifiers using the Rotation Forest method and a Maximum Probability voting classification determination approach.

With an overall F-measure of 0.728 for all seven classes—including suicidal ideation—and a specific F-measure of 0.69 for the suicidal ideation class, the study demonstrated a noteworthy accomplishment in terms of F-measure. The authors employed a 10-fold cross-validation strategy and classification metrics, such as Precision, Recall, and F-measure, to assess their classification performance

Additionally, a 12-month case study was carried out by the authors to evaluate the long-term effectiveness of their classification strategy. This case study shed light on the behaviours and demographics of Twitter users who share content linked to suicide.

As baseline classifiers, the study employed a variety of machine learning models, such as Naive Bayes (NB), Decision Trees (DT), and Support Vector Machine (SVM). Additionally, they used group techniques such as Rotation Forest (RF) to improve classification efficiency.

Comparing the automated classifier against human annotators, it attained an accuracy of 85% in the binary classification test (i.e., assessing whether an individual is suicidal). This outcome shows how well the machine learning method works to find suicidal content on Twitter.

[12] Bridianne O'Dea , Stephen Wan , Philip J. Batterham , Alison L. Calear , Cecile Paris , Helen Christensen (2015). Detecting suicidality on Twitter

In the study, researchers aimed to determine whether the content of suicide-related posts on Twitter could be reliably assessed by human coders and replicated by a machine learning classifier.

A total of 14,701 suicide-related tweets were gathered during the data collection phase. From this dataset, 14% (2,000 tweets) were randomly selected for human coding. Human coders classified these tweets into three distinct categories: 'Strongly concerning' ,'Possibly concerning' ,Safe to ignore'.

The machine learning algorithms used included Support Vector Machines (SVMs) and Logistic Regression. Different feature representations, including word frequencies and Term Frequency weighted by Inverse Document Frequency (TFIDF), were explored to represent the tweets for the classifiers. Cross-validation methods were employed to evaluate the classifier's performance, and the average accuracy when the training set was divided into 10 "folds" was assessed.
The level of agreement among the human coders was found to be 76%, with an average κ coefficient of 0.55, indicating moderate to good agreement.
Subsequently,the machine learning classifier achieved an impressive 80% accuracy in correctly identifying 'strongly concerning' tweets.
When combining both sets of human-coded data (Sets A and B), the overall accuracy of the machine classifier was 76%, demonstrating its effectiveness in replicating the accuracy of human coders.

[13] Jared Jashinsky, Scott H. Burton, Carl L. Hanson, Josh West, Christophe Giraud-Carrier, Michael D. Barnes, and Trenton Argyle (2013). Tracking Suicide Risk Factors Through Twitter in the US

The study conducted , explores the potential of Twitter as a surveillance tool to track suicide risk factors in real-time. The research aimed to identify associations between Twitter conversations containing suicide-related keywords and actual suicide rates by matching geographic data. The study used Twitter's public data via its API to collect and analyze tweets containing predefined risk factor keywords and phrases.
The study primarily used a departure from the expected ratio (dα) approach to measure the proportion of at-risk tweets in each state concerning the total tweets. States with dα values greater than 1 were considered to have a higher proportion of suicide-related tweets than expected, while values less than 1 indicated the opposite.
The study reported a Spearman's rank correlation coefficient ($r = 0.53$) to compare the Twitter-generated dα values with age-adjusted suicide rates. The correlation was statistically significant ($p < 0.001$), indicating an association between Twitter data and actual suicide rates.
The study demonstrates the potential of Twitter as a surveillance tool for tracking suicide risk factors and highlights the need for privacy safeguards and further research to validate its effectiveness in suicide prevention.

The study does not provide a specific accuracy rate, as it focuses on associations between Twitter data and actual suicide rates rather than a traditional binary classification task

[14] Patricia A. Cavazos-Rehg, Ph.D., Melissa J. Krauss, M.P.H, Shaina Sowles, M.P.H., Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J. Bierut, M.D. (2017). A content analysis of depression-related Tweets

The study analyzed a random sample of 2,000 depression-related tweets. The researchers assessed these tweets for the expression of symptoms of Major Depressive Disorder (MDD) based on the DSM-5 criteria. They also categorized the themes in these tweets and looked into the demographic characteristics of the users who posted them.

The study collected tweets that contained specific keywords related to depression, such as "depressed," "depression," and related hashtags. It gathered tweets within a specific timeframe.

The most common themes identified in the depression-related tweets were:
Supportive or helpful tweets about depression. Disclosure of feelings of depression.
Mention of school or work-related pressures related to depression. Discussion of substance use to cope with depression.

The study primarily focused on identifying common themes and symptoms in the depression-related tweets rather than traditional binary classification.

[15] Atika Mbarek , Salma Jamoussi , Anis Charfi and Abdelmajid Ben Hamadou ,(2019). Suicidal Profiles Detection in Twitter.

The study focuses on detecting suicidal profiles on Twitter, aiming to contribute to suicide prevention. With approximately 800,000 suicides occurring annually, early identification of individuals at risk is critical. The research leverages Twitter as a valuable source of data for text mining, as many suicidal individuals express their thoughts and intentions on the platform .The study combines linguistic and account-based features from Twitter profiles and tweets to classify users as either suicidal or non-suicidal. It emphasizes the importance of extracting semantic features, including linguistic, emotional, and stylometric aspects, from tweets.
The dataset comprises 115 suicidal profiles and 172 non-suicidal profiles. Supervised machine learning, implemented through classifiers like BayesNet, Adaboost, J48, SMO, and Random Forest, is employed to predict suicidal profiles. The system is implemented as a web-based Java application to assess the suicide risk of a given Twitter profile.
The study uses various tools and techniques for feature extraction and classification. Notably, the Linguistic Inquiry and Word Count (LIWC) software is employed for text analysis. Additionally, data mining tools, facial recognition APIs, and natural language processing (NLP) libraries such as

OpenNLP are used for feature extraction and analysis.
The study employs five different classifiers, including BayesNet, Adaboost, J48, SMO, and Random Forest, to classify Twitter profiles as suicidal or non-suicidal. These classifiers are evaluated for their performance in identifying suicidal profiles based on the extracted features.
The study provides precision, recall, and F-measure values for each classifier. For example, the Random Forest classifier achieved a precision of 83%, indicating its accuracy in identifying suicidal profiles. The recall rate reached 90% for both the Random Forest and Adaboost classifiers when evaluated with a testing set, highlighting the effectiveness of the method in identifying suicidal profiles.

[16] Amrita Mathur Purnima Kubde Sonali Vaidya (2020). Emotional Analysis using Twitter Data during Pandemic Situation: COVID-19

The project focuses on performing emotional analysis using Twitter data during the COVID-19 pandemic to gain insights into people's sentiments. It addresses the importance of managing mental health alongside physical health during the crisis. The goal is to provide authorities with a tool to understand public mental health and inform policies to combat the pandemic's effects on social well-being and the economy.
The project collects tweets related to COVID-19 from Twitter and performs data preprocessing, including the removal of special characters, links, and hashtags. It then applies sentiment analysis using a lexical-oriented method based on lexical databases such as the NRC emotion dictionary. The tweets are classified into basic emotions: joy, sadness, anger, fear, disgust, and surprise. Performance analysis reveals an accuracy of around 80% in classifying these emotions.

The project primarily uses R, a statistical programming language, for data analysis and sentiment classification. It also relies on data preprocessing and utilizes lexical resources like the NRC emotion dictionary. Additionally, Twitter's API and external datasets from TweetBinder are used to collect COVID-19-related tweets.
The project primarily employs a lexical-oriented method for emotion analysis based on predefined emotional keywords and dictionaries. Machine learning-based methods are not the primary approach.

The project achieves an accuracy of around 80% in classifying tweets based on basic emotions. This level of accuracy indicates the system's effectiveness in analyzing the emotional content of Twitter data during the COVID-19 pandemic and understanding the public's sentiments.

[17] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, (2017). Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python

The study reviews multiple papers on sentiment analysis concerning Twitter and introduces a generalized Python-based approach for this task. The research covers data collection, preprocessing, feature extraction, training machine learning models, and model validation. Several classifiers are evaluated, resulting in accuracy ranging from 66.24% to 90%. The study provides insights into sentiment analysis challenges and potential applications.

The study gathers Twitter data through Tweepy or employs existing data sources like Kaggle. Preprocessing steps encompass converting text to lowercase, eliminating URLs, user handles, hashtags, emoticons, and repeated characters.The research employs the TF-IDF approach for feature extraction. Various machine learning models are employed for sentiment analysis, including Support Vector Machine, Naïve Bayes, Logistic Regression, Random Forest, Neural Network, Maximum Entropy, and Ensemble classifiers. The collected dataset is partitioned into training and testing sets to train and validate the machine learning models.

Tools used were Python, NumPy, NLTK, Scikit-learn.

Models were Support Vector Machine, Naïve Bayes, Logistic Regression, Random Forest, Neural Network, Maximum Entropy, and Ensemble classifier.

Classifier uses Various machine learning classifiers are applied to evaluate sentiment in tweets.

The study reports accuracy scores for different classifiers, ranging from 66.24% to 90%. The highest accuracy achieved is 90.0% using the Maximum Entropy classifier.

[18] Riya Suchdev, Pallavi Kotkar, Rahul Ravindran, Sridhar Swamy (2014). Twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach

This research delves into Twitter sentiment analysis with a dual approach, combining knowledge-based and machine learning methodologies to effectively classify sentiments within tweets. The core objective of this study is to tackle the challenges posed by short, slang-infused, and misspelled tweets, ultimately providing valuable insights for companies seeking feedback on their products.

The primary objective of this research is to perform sentiment analysis on tweets discussing various companies, offering these companies essential feedback from a global audience.

The Sanders analytics dataset, encompassing tweets related to companies like Apple, Microsoft, and Google, serves as the fundamental data source for this study.An essential step in addressing misspellings and slang terms, as well as in removing redundant content such as emoticons and links. The study makes use of feature vectors, comprising parameters like hashtags and emoticons, to capture pertinent features from the tweets.A hybrid approach is adopted, blending knowledge-based and machine learning techniques to classify sentiments. Machine learning techniques rely on feature

vectors, while knowledge-based methodologies deal with other words, especially slang expressions.

The hybrid approach implemented in this research attains an impressive accuracy rate of 98%. This indicates the model's effectiveness in accurately categorizing sentiments in tweets.

[19] ***Yanwei Bao , Changqin Quan , Lijuan Wang , Fuji Ren(2014). The Role of Pre-processing in Twitter Sentiment Analysis .***

The principal objective envisages an attempt to study how different pre-processing strategies affect the accuracy of Twitter sentiment classification and improve it. This study takes up the Twitter Sentiment Dataset of Stanford University for analysis purposes. The worded data is subjected to sentiment analysis using a Linear classifier for the classification of tweets. This linear classifier fits the scenarios for text classification with sparse robust features. The of formal assessment is 85.5% of accuracy, exceeding the base accuracy of 80.62%.

[20] ***Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar and Shrikanth Narayanan, (2012). A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle.***

The authors developed an infrastructure that collects tweets in real-time, preprocesses the text, and uses a sentiment model to classify tweets into positive, negative, neutral, or unsure categories. The sentiment model uses a naïve Bayes classifier based on unigram features.
The system is built on the IBM InfoSphere Streams platform for real-time data processing.
Data is collected from Twitter using the Gnip Power Track, a commercial Twitter data provider.
The sentiment model was trained using Amazon Mechanical Turk annotations.
The model classifies tweets into four categories: positive, negative, neutral, or unsure.
The classifier performed at 59% accuracy for the four-category classification i.e positive, negative, neutral, or unsure.

Chapter-3

MATERIAL AND METHODS

In this comprehensive review paper, We delve into the intricacies of the past 20 papers in the field, synthesizing their findings and contributions to the domain. The primary focus is on evaluating and comparing the key performance metrics, namely F-measure, precision, and recall, employed in the various studies. By meticulously analyzing each paper's methodology, experimental setups, and reported results, I aim to provide a nuanced understanding of the advancements and challenges in the assessed approaches. This review serves as a valuable resource for researchers and practitioners seeking insights into the state-of-the-art techniques and their comparative performance in the context of F-measure, precision, and recall.

Leveraging the confusion matrix, we derived the following performance metrics using following equations:

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

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

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

Precision: Precision gauges the accuracy of positive predictions, indicating the proportion of true positives among all predicted positives. It is a crucial metric when minimizing false positives is essential.

Recall (Sensitivity): Recall measures a model's ability to capture all relevant instances by assessing the ratio of true positives to the total actual positives. It is particularly important when avoiding false negatives is a priority.

Accuracy: A metric that measures the overall correctness of a model by assessing the ratio of correctly predicted instances to the total instances.

These metrics collectively provide a comprehensive evaluation of a model's performance, balancing its ability to make accurate predictions, minimize false positives, and capture relevant instances.

where, TP—true positive, FP—false positive, TN—true negative, FN—false negative.

Precision: 0.9

Recall:0.57

F score:0.70

Accuracy:0.89

The methodology used for sentiment classification is based on the **TF-IDF (Term Frequency-Inverse Document Frequency) technique combined with a Naive Bayes classifier**.

Here's an overview of the methodology used:

1. TF-IDF Vectorization:

- The text data is preprocessed using techniques like tokenization, stop word removal, and stemming.
- The TF-IDF vectorization is applied to convert the text data into numerical feature vectors.
- TF-IDF calculates the importance of a word in a document relative to a collection of documents (corpus). It assigns higher weights to words that are frequent in a document but rare in the corpus, indicating their importance in the specific document.

2. Naive Bayes Classifier:

- After preprocessing and TF-IDF vectorization, the data is ready for classification.
- The Naive Bayes classifier is trained on the TF-IDF transformed data.
- Naive Bayes is a probabilistic classifier that calculates the probability of a document belonging to a particular class (not depressed or depressive sentiment) based on the presence of words/features in the document. It assumes independence between features, which simplifies the calculation of probabilities.

3. Prediction:

- Once the classifier is trained, it can be used to predict the sentiment of new text data.
- For each new tweet or text input, the TF-IDF vectorization is applied to transform it into a numerical feature vector.
- The Naive Bayes classifier then predicts the sentiment (positive or depressive) based on the probabilities calculated from the TF-IDF vector.

4. Average Sentiment Calculation:

- In the FastAPI application, the sentiment classification is applied to each tweet fetched from the

specified Twitter handle.

- for fetching the tweets , we are using the basic plan from twitter API developer platform.
- Instead of returning individual sentiment predictions for each tweet, the code calculates the average sentiment score based on the sentiments of all tweets fetched.
- The average sentiment score represents the overall sentiment of all tweets and is returned as a single value.

This methodology combines text preprocessing, feature engineering with TF-IDF, and classification using Naive Bayes to classify the sentiment of tweets. It provides a simple yet effective approach for sentiment analysis in this context.

Chapter-4

RESULTS AND ANALYSIS

TABLE 4.1 ANALYSIS OF MODEL USED IN DIFFERENT RESEARCH PAPERS

Research paper	Algorithm / Tools	Result
Rinku Yadav et al. (2018).	Naive Bayes Classifier	The accuracy obtained during training and testing phase is 100% and 82.2 % respectively.
Sangeeta Lal et al. (2020).	Naive Bayesian, Random Forest, J48 and ZeroR.	Random forest classifier give the best accuracy of 98.1%. Precision: 98.2 Recall : 98.1 F-measure: 98.1
Muhammad Abubakar Alhassanet al. (2021).	Latent Dirichlet Allocation (LDA) algorithm	The number of negative sentiments from the non-professional users is significantly higher on twitter.
E. Rajesh Kumaret et al.(2023).	Rotation Forest (RF) algorithm	***For suicidal ideation***: Precision: 0.764 Recall : 0.758 F-measure: 0.760 ***For all 7 classes.*** Precision: 0.813 Recall : 0.830 F-measure: 0.821
Selva Mary G. et al. (2023).	Integrating word2vec and BERT with Bi-LSTM	achieved a remarkable 98.5% accuracy
Krishna Shrestha (2018).	Review previous studies that employed machine learning for identifying depression.	Observed Naïve Bayes Linear SVM Logistic Regression Decision Tree has highest accuracy of Accuracy = 86%
Michael Mesfin Tadesse et al. (2019)	Tried to demonstrate that LSTM-CNN model can outperform the performance of its individual CNN and LSTM.	Accuracy: 93.8% Precision: 93.2% Recall : 94.1% F1 Score: 93.4%
Shaoxiong Ji et al.(2018)	compare six classifiers, including four traditional supervised classifiers and two neural network models including Random Forest, GBDT, XGBoost, SVM, MLFFNN and LSTM	RF gives highest accuracy of 96%. Precision: 0.9638 Recall: 0.9917 F1 score: 0.9646
Minsu Park et al.(2012).	Conducted a study on 69 participants to determine whether the use of sentiment words	Based on the CES-D cutoff score of 22, 41 participants were classified into low or mild depression and 28 participants scored positive for depression.

	of depressed users differed from normal user.	
Qurat Tul Ain et al. (2017).	The review has described studies related to sentiment analysis by using deep learning models.	The capability of settling in the task variations by having little alterations in system itself includes a feather in strength of Deep Learning.
Pete Burnapa et al.(2017).	Rotation Forest method and a Maximum Probability voting classification, determination approach such as Naive Bayes (NB), Decision Trees (DT), and Support Vector Machine (SVM).	***For all 7 classes.*** F-measure :0.728 **accuracy** of 85% in the binary classification test
Bridianne O'Dea et al.(2015).	Support Vector Machines (SVMs) and Logistic Regression, frequencies and Term Frequency weighted by Inverse Document Frequency (TFIDF), Cross-validation methods	human coders **accuracy** 76%, **average κ coefficient** :0.55 machine learning **accuracy** 80%
Jared Jashinsky et al. (2013).	expected ratio (dα) approach, the study reported a Spearman's rank correlation coefficient (r = 0.53) to compare the Twitter-generated dα values with age-adjusted suicide rates.	The study does not provide a specific accuracy rate, as it focuses on associations between Twitter data and actual suicide rates
Patricia A. Cavazos-Rehg et al. (2017).	Demographics Pro tool to infer demographic characteristics of Twitter users who posted depression-related content	It primarily focused on identifying common themes and symptoms in the depression-related tweets
Atika Mbarek et al.(2019)	Supervised machine learning, classifiers like BayesNet, Adaboost, J48, SMO, and Random Forest.	Precision : 83% Recall : 90%
Amrita Mathur et al.(2020).	lexical-oriented method	Accuracy : 80%
Bhumika Gupta et al. (2017).	Support Vector Machine, Naïve Bayes, Logistic Regression, Random Forest, Neural Network, Maximum Entropy, and Ensemble classifiers	Accuracy ranging from 66.24% to 90%. The highest accuracy achieved is 90.0% using the Maximum Entropy classifier.

Riya Suchdev et al. (2014).	**Knowledge-based:** bag-of-words, lexical database WordNet, word space model formalism, EmotiNet, coarse-grained and fine-grained approaches **Machine learning:** Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) a	The **hybrid approach** implemented in this research attains an **impressive accuracy rate of 100%**
Yanwei Bao et al. (2014).	Liblinear classifier (tool)	Sentiment classification accuracy of 85.5%
Hao Wang, et al. (2012).	Naïve Bayes classifier	Classifier performed at 59% accuracy.

After API integration response and output

For average:

Input: user_handle

Output: average sentiment of 200 retrieved tweets to classify the person as depressive or not


```
{
  "user_handle": "ElonMusk",
  "sentiment": "not depressed"
}
```

Chapter-5

DISCUSSION AND CONCLUTION

This comprehensive review delves into the intricate landscape of sentiment analysis within the context of social media platforms, primarily focusing on Twitter. The explored research papers collectively contribute valuable insights into diverse realms, including the detection of self-harm tendencies, analysis of crime-related tweets, and understanding discussions related to non-suicidal self-injury.

The methodologies employed in these studies span a spectrum of approaches, ranging from traditional machine learning algorithms like Naive Bayes and Random Forest to advanced techniques such as deep learning models and sentiment analysis tools. The effectiveness of these methodologies is evident in their high accuracy rates, ranging from 82.2% to 98.1%, showcasing the robustness of the proposed solutions in addressing complex challenges posed by online content.

Notably, the papers highlight the significance of sentiment analysis not only for understanding user emotions and tendencies but also for addressing critical issues such as mental health awareness, crime management, and suicide prevention. The interdisciplinary nature of these studies underscores the growing importance of leveraging social media data to enhance public safety, mental health support, and overall well-being.

As we navigate the evolving digital landscape, the findings from these research papers serve as a foundation for future endeavors in developing proactive measures, improving classification accuracy, and extending these methodologies to other social media platforms. The collective knowledge synthesized in this review emphasizes the pivotal role of sentiment analysis in creating safer and more supportive online environments, ultimately contributing to the broader goal of fostering a responsible and empathetic digital realm.

Chapter-6

SUMMARY, PUBLICATIONS AND FUTURE WORK

Student will write their own chapter titles, topics, sub topics, sub sub topics. Font is Times New Romans, Size 12, Spacing 1.5. This chapter will give the summary of what were the objectives and how much you succeed in achieving them preferably with some facts and figures or percentages of errors/ accuracy or summarized numerical justification of the parameters used in the project [1-5] . nonummy sagittis nonummy posuere sed et habitant vehicula leo odio ultricies fermentum felis dui blandit aptent vitae id et diam vitae molestie, aenean, porttitor mattis, taciti tincidunt sodales massa vulputate fames scelerisque sollicitudin. Nunc viverra Ultrices placerat. Platea taciti. Gravida adipiscing mattis Proin commodo morbi sed consequat. Libero blandit. Student will write their own chapter titles, topics, sub topics, sub sub topics.

6.1 FUTURE WORK

Following is the sample of All heading lev There is still a great deal of work to be done, but we hope to shed some light on potential study directions in the future.

- *Interpreting Sarcasm*:At the moment, the suggested method is unable to understand sarcasm. Sarcasm often refers to the use of irony to belittle or express dislike; in the context of contemporary art, sarcasm is the conversion of a supposedly positive or negative statement into its inverse.

- *Multimodal Sentiment Analysis*: Future work can explore the integration of multiple data modalities, including text, images, and possibly audio, to achieve a more comprehensive understanding of user sentiment. This can enhance the accuracy of sentiment analysis models by capturing nuanced emotional expressions that may be conveyed through diverse media.

- *Real-time and Streaming Analysis*: Given the dynamic nature of social media, developing sentiment analysis models capable of real-time and streaming analysis is crucial. This would enable timely interventions in situations related to mental health crises or emergencies, enhancing the effectiveness of support systems.

- *Cross-Platform Generalization*: Extending sentiment analysis models to encompass multiple social media platforms beyond Twitter could be a fruitful area of exploration. Each

platform has its unique characteristics, and devising models that generalize well across platforms would broaden the impact of sentiment analysis in safeguarding user well-being.

- *Human-AI Collaboration*: Exploring ways to integrate human judgment into sentiment analysis systems can lead to hybrid models that leverage the strengths of both AI algorithms and human intuition. Such collaborative approaches may improve the accuracy and relevance of sentiment classifications.

By addressing these future directions, researchers can further advance the field of sentiment analysis, making it more robust, ethical, and attuned to the evolving needs of users in the digital age. This ongoing exploration will contribute to the development of safer, more supportive online environments and enhance the practical applications of sentiment analysis across diverse domain sets.