

CLASSIFICATION OF DEPRESSION ON SOCIAL MEDIA USING TEXT MINING



Department of Computer Science and Engineering

Netaji Subhas University of Technology (NSUT)

NAME OF THE STUDENT

ROLL NUMBER

ADITYA BABU

2020UCM2318

FARIS REHMAN

2020UCM2341

DIVYANSH SAGAR

2020UCM2343

Under the supervision

Of

PROF. SANGEETA SABHARWAL

Department Of Computer Science Engineering
Netaji Subhas University of Technology, New Delhi



Certificate

This is to certify that the work embodied in the project thesis titled “**Classification of Depression on Social Media Using Text Mining**” by Aditya Babu (2020UCM2318), Faris Rehman (2020UCM2341) and Divyansh Sagar (2020UCM2343) is the bonafide work of the group submitted to **Netaji Subhas University Of Technology** for consideration in 7th Semester B. Tech. Project Evaluation.

The original work was carried out by the team under my/our guidance and supervision in the academic year 2023-2024. This work has not been submitted for any other diploma or degree from any university.

On the basis of a declaration made by the group, we recommend the project report for evaluation.

Dr. Sangeeta Sabharwal

(Professor)

Department of Computer Science and Engineering
Netaji Subhas University of Technology

Department Of Computer Science Engineering
Netaji Subhas University of Technology, New Delhi



CANDIDATE(S) DECLARATION

I/We, Aditya Babu (2020UCM2318), Faris Rehman (2020UCM2341) and Divyansh Sagar (2020UCM2343) of B. Tech. Department of Computer Science & Engineering(Mathematics and Computing), hereby declare that the Project titled “Classification of Depression on Social Media using Text Mining” which is submitted by me/us to the Department of Computer Science & Engineering, Netaji Subhas University of Technology (NSUT) Dwarka, New Delhi in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology is original and not copied from the source without proper citation. The manuscript has been subjected to plagiarism checks by Turnitin software. This work has not previously formed the basis for the award of any degree.

Place: Dwarka, New Delhi
Date: 10/12/2023


Aditya Babu

2020UCM2318


Faris Rehman

2020UCM2341



Divyansh Sagar

2020UCM2343

ACKNOWLEDGEMENT

We would like to express our gratitude and appreciation to all those who made it possible to complete this project. Special thanks to our project supervisor **Dr. Sangeeta Sabharwal** help, stimulating suggestions and encouragement helped us in writing this report. We also sincerely thank our colleagues for the time spent proofreading and correcting our mistakes.

We would also like to acknowledge with much appreciation the crucial role of the staff in Computer Science and Engineering, who gave us permission to use the lab and the systems and gave permission to use all necessary things related to the Project.



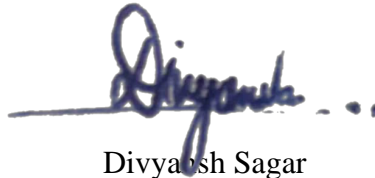
Aditya Babu

2020UCM2318



Faris Rehman

2020UCM2341



Divyansh Sagar

2020UCM2343

ABSTRACT

In an era dominated by digital communication, the escalating prevalence of mental health issues, particularly depression, necessitates innovative solutions. This B. Tech project serves as a beacon of hope and progress, acknowledging the global reach of mental health challenges. Our shared vision involves harnessing the vast repository of social media data, encompassing diverse human experiences shared through platforms such as tweets and Facebook posts, as a potent instrument for positive change.

By applying advanced machine learning techniques, our research strives to unravel the intricacies of mental health, identifying patterns and insights that can inform evidence based interventions, shape enlightened public policies, and fortify support systems. Delving into the multifaceted challenges faced by individuals grappling with depression in the digital age, our objective extends beyond raising awareness to inspire actionable solutions. Ultimately, we aspire to forge a lasting, tangible impact on the lives of those affected by depression, envisioning a world where mental well-being is universally valued, and no one suffers in silence.

TABLE OF CONTENT

Sr. No.	Description	Page Number
1.	Certificate	1
2.	Candidates Declaration	2
3.	Acknowledgement	4
4.	Abstract	5
5.	Chapter 1 - Introduction	9
6.	1.1 - Motivation	10 - 11
7.	1.2 - Literature Review	11 - 13
8.	1.2 - Key Challenges	13 - 15
9.	1.3 - Problem addresses in thesis	15 - 16
10.	1.4 - Approach to the problem and organization of the thesis	16 - 19
11.	Chapter 2 - Mathematical Model/Experimentation Methods And Materials	20
12.	2.1 - Overall Tools and Technologies Used	20
13.	2.2 - Libraries/ Dependencies of the Project	20 - 21
14.	2.3 - Methodology	22 - 23
15.	Chapter 3 - Implementation	24
16.	3.1 - Implementation steps	24 - 25
17.	3.2 - Some more snapshots and workflows	25 - 26
18.	3.3 - Result from different classifiers more elaborated	27 - 28
19.	3.4 - Sample tweets more elaborated	28 - 31
20.	Chapter 4 - Conclusion and scope for future work	32
21.	4.1 - Task achievement, and possible beneficiaries	32 - 33

Sr. No.	Description	Page Number
22.	4.2 - Review of contributions	33 - 34
23.	4.3 - Scope for future work	35
24.	References	36
25.	Plagiarism Report	37

TABLE OF FIGURE

Sr. No.	Description	Page Number
1	Fig 2.3.1 - Test Data Set Diagram	22
2	Fig 3.2.1 – Dictionary Snapshot	25
3	Fig 3.2.2 – Tweet Data Snapshot	26
4	Fig 3.2.3 – Output Snapshot	26
5	Fig 3.4.1 – Processing	29
6	Fig 3.4.2 – Preprocessing Done Figure	30
7	Fig 3.4.3 – Sample Tweet	30
8	Fig 3.4.4 – Accuracy Check And Completing Time	31
9	Fig 3.4.5 – Negative Tweet Demo	31
10	Fig 3.4.6 – Neutral Tweet Demo	31
11	Fig 3.4.7 – Positive Tweet Demo	31

CHAPTER 1

Introduction

In the contemporary era, where social media platforms like Twitter and Facebook serve as pervasive modes of communication, ensuring the mental well-being of users has gained paramount importance. The prevalence of mental health issues, particularly depression, within the digital landscape underscores the critical need for innovative solutions. This project, centered on the classification of depression using text mining on social media, aims to contribute to the enhancement of mental health awareness and support in the digital sphere.

Given the escalating frequency of mental health-related discussions on platforms like Twitter and Facebook, the imperative for robust solutions becomes evident. Through the application of advanced text mining techniques, this B.Tech project seeks to develop a comprehensive framework for identifying and categorizing expressions indicative of depression in social media text.

The fundamental approach involves leveraging state-of-the-art text mining methodologies to extract meaningful insights from user-generated content. By adopting a nuanced understanding of linguistic patterns, sentiment analysis, and contextual cues, the project aspires to enhance the accuracy of depression classification. This involves addressing the inherent challenges of noise, ambiguity, and context-specific language prevalent in social media discourse.

Similar to the integration of traditional mail systems with encryption technologies in the email security project, our initiative acknowledges the dynamic nature of social media communication. The aim is to seamlessly integrate text mining methodologies with the existing fabric of platforms like Twitter and Facebook. This user-centric approach is pivotal, emphasizing both the accuracy of depression classification and the overall user experience.

In crafting a comprehensive solution that amalgamates cutting-edge text mining methods with widely used social media platforms, this project endeavors to make a substantive impact on the intersection of mental health awareness and digital communication. The adoption of this classification system is positioned to offer timely and creative responses to the evolving challenges of identifying and supporting individuals experiencing depression within the digital realm. Recognizing that social media remains a vital channel for interpersonal and professional interactions, the project strives to contribute meaningfully to the well-being of users navigating the complexities of mental health in the digital age.

1.1 Motivation

In the contemporary digital age, social media has emerged as a pervasive platform for communication, providing individuals with an unprecedented means to express their thoughts and emotions. However, this vast virtual space is not devoid of challenges, and one pressing concern is the escalating issue of mental health, particularly the prevalence of depression. As we witness an increasing reliance on social media for personal expression, the need to identify and support individuals experiencing mental health challenges becomes paramount.

The motivation behind this project stems from a profound recognition of the profound impact social media can have on mental health and the imperative to leverage technology for positive change. The ubiquity of social media platforms makes them an invaluable source of real-time data, offering a unique opportunity to detect signs of depression through the analysis of textual content.

The landscape of mental health is complex, and traditional methods of identification often fall short in the dynamic and evolving online environment. By harnessing the power of text mining and sentiment analysis on social media data, this project aims to contribute to the early detection and classification of depression, offering a potential avenue for timely intervention and support.

The motivation to undertake this project is not merely technological but driven by a genuine concern for the well-being of individuals within our digital society. Depression, often hidden behind the veil of online communication, presents a challenge that demands innovative solutions. By delving into the linguistic nuances of social media posts, we aspire to create a tool that goes beyond the surface, providing insights into the emotional well

being of users and fostering a supportive online community.

The significance of addressing mental health challenges on social media extends beyond the individual to societal well-being. As we navigate the complexities of the digital age, the successful implementation of this project holds the promise of not only identifying those in need but also paving the way for empathetic and targeted interventions, ultimately contributing to a healthier and more compassionate online environment.

In summary, the motivation behind this project is driven by a commitment to leveraging technology for the betterment of mental health, recognizing the unique challenges posed by social media, and aiming to make a positive impact on individuals' lives within the digital realm.

the expansive realm of social media. The motivation to delve into this project is rooted not just in technological innovation but in a sincere commitment to creating a more compassionate and supportive online environment.

The digital landscape, with its vast troves of user-generated content, serves as a rich source

of information that extends beyond mere words. This project recognizes the nuances and intricacies of human expression on social media—subtle changes in language, shifts in sentiment, and patterns of interaction that may signal distress. By adopting a text mining approach, we aim to unravel these complexities and contribute to a proactive system capable of identifying and classifying potential instances of depression.

The decision to focus on social media as a domain for mental health exploration is driven by the sheer prevalence of online interactions. People often turn to these platforms to share their thoughts, seek solace, or connect with others. Consequently, the digital footprint left behind becomes a valuable reservoir of data that, when analyzed judiciously, could offer unprecedented insights into the mental well-being of individuals.

Moreover, the motivation extends beyond the academic realm and into the heart of societal welfare. Depression, if left unaddressed, not only takes a toll on individual lives but also has cascading effects on communities and societies. By creating a tool that facilitates the early detection of depression in the digital sphere, this project aspires to contribute to the broader discourse on mental health, fostering a collective responsibility for the well-being of online communities.

❖ 1.2 Literature Review

➤ **Deep Learning for Depression Detection from Textual Data [fMRI Signature [19]]:**

- The paper titled "Deep Learning for Depression Detection from Textual Data" by Amna Amanat, Muhammad Rizwan, Abdul Rehman Javed, Maha Abdelhaq, Raed Alsaqour, Sharnil Pandya, and Mueen Uddin explores the use of deep learning techniques for detecting depression from textual data. The paper likely explores the potential of deep learning in leveraging textual data for depression detection. While offering promising insights, challenges such as data limitations, interpretability, ethical considerations, and cross-cultural applicability should be addressed for the successful implementation of such methods in real-world scenarios.

◆ **Salient Features:**

Focuses on utilizing deep learning techniques for depression detection from textual data.

Likely involves natural language processing (NLP) methods.

◆ **Shortcomings:**

Dependency on the availability and quality of textual data.

Potential challenges in interpreting nuanced language and context.
Ethical concerns regarding privacy and consent in handling sensitive textual information.

➤ **Studies using Depression Questionnaires (DASS21 and DASS42) [1,20]:**

- The paper "Deep Learning for Depression Detection from Textual Data" explores the application of deep learning techniques, likely involving neural networks and natural language processing, to analyze textual data for detecting depression. While promising, the study may face challenges such as limited data quality, interpretability issues with deep learning models, ethical considerations in handling sensitive information, and potential variations in performance across different cultural and linguistic contexts. The research contributes to improving depression detection in mental health but needs careful consideration of these challenges for practical implementation.

- ◆ **Salient Features:**

- Relies on self-reported data from individuals through standardized depression questionnaires.

- Enables quantitative assessment of various dimensions of depression.

- ◆ **Shortcomings:**

- Susceptible to bias due to self-reporting inaccuracies.

- May not capture the full spectrum of an individual's mental health.

- Dependence on participants' willingness to disclose information truthfully.

➤ **Utilizing Clinical Criteria from DSM-5 and ICD-10 [21]:**

- The paper investigates using deep learning, including neural networks and natural language processing, for detecting depression from textual data. It shows promise in mental health applications but faces challenges like limited data quality, interpretability issues with deep learning models, ethical concerns, and potential variations in performance across different cultural contexts. The study contributes to enhancing depression detection, emphasizing the need to address these challenges for practical implementation.

- ◆ **Salient Features:**

Applies standardized clinical criteria for depression diagnosis.
Aligns with established medical and psychological guidelines.

◆ **Shortcomings:**

Diagnosis might rely on subjective judgment, leading to potential variability.

Dependent on the accuracy of the initial clinical assessment.

May not be suitable for real-time or continuous monitoring.

1.3 Key Challenges

Addressing the nuanced landscape of depression classification through text mining on social media platforms such as Twitter and Facebook presents a set of intricate challenges. As a final year engineering student, it is essential to delineate and comprehend these challenges that will shape the trajectory of the project.

❖ **Linguistic Complexity and Contextual Understanding:**

Analyzing social media content for signs of depression involves navigating linguistic complexities and context-specific expressions. The challenge lies in developing algorithms that can comprehend the subtleties of language, slang, and cultural nuances to accurately identify and classify depressive content.

❖ **Data Privacy and Ethical Considerations:**

The project must grapple with the challenge of handling sensitive user data ethically. Balancing the need for data to train robust classification models with respect for user privacy requires the implementation of stringent data anonymization techniques and adherence to ethical guidelines.

❖ **Imbalanced Data Sets:**

Social media data often exhibits class imbalance, where depressive content may be significantly outnumbered by non-depressive content. Addressing this challenge involves implementing strategies such as oversampling, undersampling, or utilizing advanced machine learning techniques to ensure the model is not biased towards the majority class.

❖ **Dynamic Nature of Language and Trending Topics:**

Social media content is dynamic, reflecting real-time conversations and evolving trends. The challenge lies in creating a system that can adapt to the ever-changing nature of language and identify emerging expressions related to depression, ensuring the model's

relevance over time.

❖ **Cross-Platform Variability:**

Twitter, Facebook, and other platforms exhibit variations in user behavior and communication styles. Adapting the classification model to account for these differences poses a challenge, requiring the development of platform-specific models or a unified approach that accommodates diverse linguistic patterns.

❖ **Validation and Ground Truth Challenges:**

Establishing a robust ground truth for training and validating the classification model is challenging. Labeling social media posts as depressive or non-depressive involves subjective judgment, and obtaining a reliable ground truth dataset is crucial for the model's accuracy.

❖ **User Engagement and Trust:**

Convincing users to engage with the classification system and share their social media data for research purposes presents a challenge. Building trust through transparent communication about the project's objectives and ensuring user consent is crucial for the success of the endeavor.

❖ **Interpretability and Explain-ability:**

Developing a model that not only accurately classifies depression but also provides interpretable results is a challenge. Ensuring that users, researchers, and mental health professionals can understand and trust the decisions made by the model is essential for widespread acceptance.

❖ **Continuous Model Improvement:**

The dynamic nature of social media language requires a continuous effort in refining and improving the classification model. Implementing mechanisms for ongoing model training and adaptation to evolving language trends is a persistent challenge.

❖ **Ethical Implications and Bias Mitigation:**

Detecting and mitigating biases in the classification model is crucial to avoid perpetuating stereotypes or causing harm. The challenge is to implement fairness-aware machine learning techniques and ensure that the model's predictions do not disproportionately affect certain demographic groups.

In conclusion, the challenges in classifying depression on social media through text mining

are multifaceted, encompassing linguistic nuances, ethical considerations, data imbalance, and the dynamic nature of online communication. Addressing these challenges requires a meticulous and ethical approach, blending advanced machine learning techniques with a deep understanding of the social and cultural context within which the classification occurs. As a final year engineering student, navigating these challenges offers an opportunity for meaningful contributions to the intersection of technology and mental health.

1.4 Problem Addressed in Project

This project directs its focus towards the pivotal challenge of advancing mental health awareness and support in the digital age, specifically in the realm of social media platforms such as Twitter and Facebook. As these platforms have become integral to personal expression and communication, the burgeoning concern of identifying and addressing depression in this dynamic digital space demands meticulous attention.

❖ Prevalence of Mental Health Expression on Social Media:

With social media serving as a prevalent outlet for personal expression, the challenges of identifying and addressing mental health issues, particularly depression, have surfaced prominently. The project acknowledges the need to navigate this expansive digital landscape to discern signs of depression amidst diverse user interactions.

❖ Limitations in Current Mental Health Identification Practices:

Current methodologies for identifying and addressing mental health concerns on social media often encounter limitations due to the dynamic nature of language and evolving expressions of users. This project specifically aims to overcome these limitations and develop a more effective and robust approach to classify depression through text mining.

❖ Integration of Advanced Text Mining Techniques:

Similar to the integration of cryptographic techniques in the email security thesis, this project strategically incorporates advanced text mining techniques to classify depression. The goal is to establish a comprehensive framework that is both theoretically grounded and practically applicable in the realm of social media.

❖ User-Friendly Integration with Social Media Platforms:

The overarching problem addressed is the imperative to devise a system that not only accurately classifies depression but also ensures user-friendly integration with popular social media platforms. The project aims to create a solution that is accessible to end-users

while effectively identifying and addressing mental health concerns.

❖ **Mitigation of Current Vulnerabilities in Mental Health Practices:**

In addition to the primary focus on classification, the project endeavors to mitigate vulnerabilities present in current mental health identification practices on social media. By leveraging advanced text mining techniques, the aim is to enhance the accuracy and sensitivity of the system in identifying potential cases of depression.

❖ **Real-world Challenges Faced by Individuals and Organizations:**

The research emphasizes the practical implementation of the proposed text mining system to address real-world challenges faced by individuals and organizations navigating the complexities of mental health on social media. This includes considerations of privacy, ethical implications, and the need for timely and targeted interventions.

In conclusion, the problem addressed in this project is intricate, involving the enhancement of accuracy in depression classification on social media, integration of advanced text mining techniques, and creation of a user-friendly solution aligned with practical realities of social media usage. This research aims to contribute meaningfully to the ongoing discourse on mental health and technology, offering a comprehensive and widely adoptable solution tailored for the digital age.

1.5 Approach to the Problem and Organization of the Thesis

❖ **Approach to the Problem:**

Addressing the intricate challenge of classifying depression through text mining on social media involves a comprehensive approach that integrates advanced natural language processing techniques, user-centric design principles, and a proactive stance towards emerging mental health concerns. The approach adopts a systematic and iterative methodology to develop a classification system that leverages the unique characteristics of text data on platforms like Twitter and Facebook.

❖ **Understanding Current Challenges:**

The initial phase involves a meticulous analysis of the current challenges in identifying and classifying depression on social media. This includes an in-depth examination of existing text mining techniques, sentiment analysis limitations, ethical considerations, and the evolving nature of mental health expressions in online spaces.

❖ **Selection of Text Mining Techniques:**

The choice of specific text mining techniques is informed by their applicability to social media data. Natural language processing algorithms, sentiment analysis, and machine learning models are explored for their potential in discerning indicators of depression from user-generated content.

❖ **Classification Framework:**

The thesis proposes the development of a classification framework that seamlessly integrates various text mining techniques. This involves designing algorithms for sentiment analysis, topic modeling, and context-aware classification. The goal is to optimize the accuracy of depression classification while considering the dynamic nature of online language.

❖ **User-Centric Design:**

Recognizing the sensitivity of mental health issues, the thesis emphasizes a user-centric design approach. This involves iterative design processes, usability testing, and feedback loops to refine the user interface, ensuring that the classification system is intuitive, respectful, and supportive for users expressing mental health concerns.

❖ **Interoperability Solutions:**

To address the diverse nature of social media platforms, the thesis outlines strategies for ensuring interoperability. Compatibility with popular platforms like Twitter and Facebook is prioritized to create a classification system that can adapt to the varied linguistic expressions and communication styles across different platforms.

❖ **Security and Ethical Considerations:**

The proposed classification system undergoes a thorough examination of security and ethical considerations. This includes privacy safeguards, ensuring user consent, and preventing the misuse of mental health information. The aim is to create a system that prioritizes user well-being while maintaining ethical standards.

❖ **Testing and Validation:**

The proposed classification system undergoes rigorous testing and validation. This includes assessing the system's accuracy, sensitivity, and specificity in identifying depression indicators. User feedback and real-world testing contribute to refining the system for optimal performance.

❖ **Scalability and Adaptability:**

Recognizing the potential growth in user-generated content, the thesis outlines measures to ensure the system can scale seamlessly. This involves optimizing the system's performance and resource usage to accommodate varying computational capabilities and storage constraints, ensuring scalability as the user base expands.

❖ **Ethical Considerations:**

The thesis addresses ethical considerations by ensuring the classification system aligns with data protection and privacy laws. This involves a comprehensive understanding of regional and international regulations to ensure compliance while preserving functionality and user privacy.

❖ **Continuous Improvement:**

Given the dynamic nature of mental health expressions and online communication, the thesis emphasizes the need for continuous improvement. Regular updates and refinements to the classification algorithms are envisioned to adapt to evolving linguistic trends and enhance the system's effectiveness over time.

❖ **Organization of the Thesis:**

The thesis is structured to provide a coherent and logical presentation of the research journey. The organization is as follows:

Introduction:

- Provides an overview of the significance of classifying depression on social media.
- Outlines the objectives, research questions, and the relevance of the study.

Literature Review:

- Surveys existing literature on text mining, sentiment analysis, and mental health classification on social media.
- Identifies gaps and limitations in current approaches.

Current Challenges in Depression Classification:

- Explores the complexities and challenges associated with identifying depression through text mining on social media.

Proposed Classification Framework:

- Details the design and development of the classification framework, incorporating text mining techniques and user-centric design principles.

Implementation and Methodology:

- Describes the practical steps taken to implement the classification system.
- Outlines the methodologies for data collection, preprocessing, and model training.

Results and Validation:

- Presents the results of testing and validation, including the system's accuracy, sensitivity, and specificity.
- Discusses user feedback and real-world testing outcomes.

Ethical Considerations and Privacy Safeguards:

- Addresses ethical considerations and privacy safeguards implemented in the classification system.

Scalability and Adaptability:

- Discusses measures taken to ensure the scalability and adaptability of the classification system.

Conclusion:

- Summarizes key findings.
- Discusses the implications of the research and potential future developments.

References:

- Lists all sources cited in the thesis.

Appendices:

- Includes supplementary materials such as code snippets, additional data, or detailed information on specific methodologies.

Chapter 2

Mathematical Model/Experimentation Methods And Materials

2.1 Overall tools and technologies used

Python 3.6.1 or Higher:

- ❖ Python:
 - Python is a high-level, general-purpose programming language known for its readability and versatility. In this project, Python served as the foundational language for coding. The specified version, 3.6.1 or higher, indicates the compatibility requirements with specific libraries and modules used in the project.

Twitter Developer Account:

- ❖ Twitter API:
 - The Twitter Developer Account was crucial for accessing the Twitter API. This account provided essential credentials:
 - Consumer Key: A unique key for the application.
 - Consumer Secret: A secret key paired with the consumer key.
 - Access Token: Grants access to a user's Twitter account.
 - Access Secret: Paired with the access token for secure authentication.
 - These credentials were necessary to retrieve tweets related to depression from the Twitter platform.

These two components formed the foundational tools for data retrieval and programming in the project, ensuring compatibility and access to the required data sources.

2.2 Libraries/Dependencies of the project

The successful execution of the project relied on several essential libraries and dependencies in the Python ecosystem. These libraries provided specialized functionalities, ranging from handling data to implementing machine learning algorithms.

- ❖ Keras:
 - Keras is an open-source high-level neural networks API, written in Python and capable of running on top of TensorFlow. It simplifies the process of building and

experimenting with artificial neural networks, contributing to the project's machine learning aspects.

❖ **TensorFlow (TF):**

- TensorFlow is an open-source machine learning library developed by the Google Brain team. It is widely used for various machine learning and deep learning applications. In this project, TensorFlow served as a core library for implementing and training machine learning models.

❖ **NumPy:**

- NumPy is a powerful library for numerical operations in Python. It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions. NumPy played a crucial role in handling and manipulating numerical data efficiently.

❖ **scikit-learn (sklearn):**

- Scikit-learn is a machine learning library that provides simple and efficient tools for data analysis and modeling. It includes various classification, regression, clustering, and dimensionality reduction algorithms. Sklearn was utilized for building and evaluating machine learning classifiers.

❖ **Pandas:**

- Pandas is a data manipulation and analysis library. It offers data structures for efficiently manipulating large datasets and tools for data cleaning, transformation, and analysis. In this project, Pandas facilitated preprocessing and organizing the dataset.

❖ **itertools:**

- itertools is a module in the Python Standard Library that provides fast, memory-efficient tools for handling iterators. It was likely used for efficient looping and iteration purposes within the project.

These libraries collectively formed the backbone of the project, enabling the implementation of machine learning models, efficient data handling, and overall project functionality.

2.3 METHODOLOGY

Prerequisites:

- ❖ Utilize Python version 3.6.1 or above.
- ❖ Possess a Twitter developer account.
- ❖ Incorporate essential modules, including but not limited to Keras, TensorFlow (TF), NumPy, Scikit-learn (SKlearn), and Pandas.
- ❖ Cultivate a substantial level of patience and foster a genuine passion for machine learning.

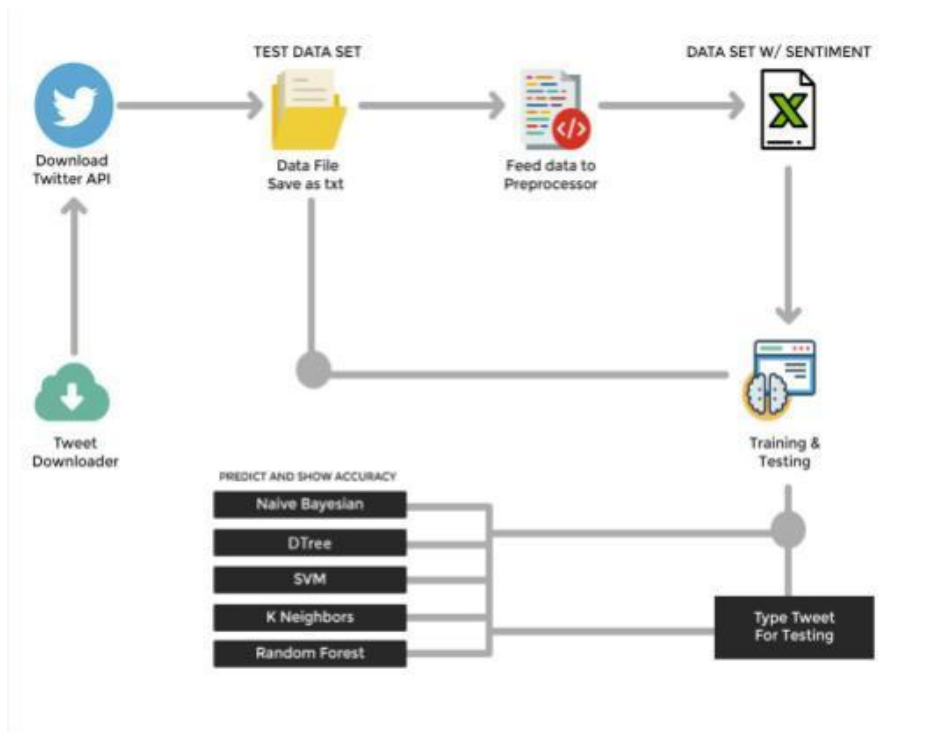


Fig 2.3.1 - Test Data Set Diagram

❖ **Test Data Set Diagram:**

- The visual representation encapsulates the holistic journey of a test data set within a machine learning ecosystem, offering a clear and intuitive flowchart.

❖ **Twitter API Integration:**

- Commencing the process, the test data set embarks on a transformative journey as it interfaces with the Twitter API. This interaction results in the retrieval and subsequent storage of the data, captured elegantly in the diagram.

❖ **Text File Preprocessing:**

- The journey takes a crucial turn as the stored text file undergoes meticulous preprocessing. This stage involves the refinement of textual content by eliminating stop words, applying word stemming, and converting the textual information into numerical features – a crucial step in preparing the data for machine learning algorithms.

❖ **Machine Learning Algorithm Application:**

- The image portrays versatility in model selection, showcasing a diverse ensemble of machine learning algorithms. From Naive Bayes to Decision Trees, Support Vector Machines, K-Nearest Neighbors, and Random Forest, the array of models reflects the richness and flexibility in algorithmic choices.

❖ **Training and Testing:**

- The heart of the process lies in the training phase, where the algorithms absorb insights from the preprocessed data. Subsequently, the models undergo rigorous testing on an independent test set, a pivotal step to gauge their predictive prowess and generalization capabilities.

❖ **Performance Evaluation:**

- The image doesn't merely stop at model execution; it ventures into the realm of performance evaluation. Beyond the traditional metric of accuracy, it delves into precision, recall, and the F1 score, demonstrating a nuanced understanding of model assessment that extends beyond a single metric.

Additional Points:

❖ **Diverse Test Data Sets:**

- Acknowledging the multifaceted nature of data, the image emphasizes the potential diversity in test data sets. It acknowledges that these sets can take varied forms, serving distinct purposes such as classification or regression.

❖ **Versatile Data Preprocessing Techniques:**

- Flexibility is a key theme as the image recognizes the varied preprocessing techniques available. From the removal of stop words to word stemming and the transformation of text into numerical features, the diagram acknowledges the adaptability required in preparing data for different machine learning contexts.

CHAPTER 3

IMPLEMENTATION, RESULTS AND DISCUSSION

3.1 IMPLEMENTATION STEPS

❖ Establish a Twitter Developer Account:

Initiate the creation of an account on the Twitter Developer platform, a prerequisite for accessing the Twitter API. Extract essential credentials—consumer_key, consumer_secret, access_token, and access_secret—from this account, pivotal for authentication and programmatic access to Twitter data.

❖ Retrieve Tweets Using Twitter API:

Execute the "Download_twitter_Api.py" script, inserting the obtained credentials. This script facilitates the retrieval of current tweets containing specific keywords like depression, anxiety, or sadness. The acquired datasets will serve as the foundational data for subsequent analysis.

❖ Preprocessing Phase - Execute "preprocessor.py":

Prepare the acquired data for analysis by running the "preprocessor.py" script. This stage involves navigating through datasets and a predefined dictionary, associating words with their respective polarities—a crucial element in determining tweet sentiment. The sentiment of each tweet is computed by summing up individual word polarities and dividing by the total number of words in the tweet.

❖ Processed Data Output:

Store the preprocessed data in the "processed_data/output.xlsx" directory. Upon opening this Excel file, you'll find tweet IDs and their corresponding sentiments (Positive, Neutral, Negative) in separate columns. This output represents a filtered Twitter dataset based on depressing keywords.

❖ Training and Prediction - Execute "depression_sentiment_analysis.py":

Ensure correct file locations and run the "depression_sentiment_analysis.py" script. This code processes the "output.xlsx" file, retrieves original tweets corresponding to sentiment

IDs, and feeds the data to classifiers. The console output will display the Area Under the Curve (AUC) for each classifier, offering a performance measure. Interpreting Results:

- ❖ **Positive:** Signals a likelihood that the individual does not currently exhibit signs of depression or anxiety based on the content of their tweets.
- ❖ **Neutral:** Implies a middle ground where the user's mental state is uncertain, with the possibility of experiencing depression or displaying depression-like symptoms.
- ❖ **Negative:** Indicates a lower level where evident signs of depression and anxiety are identified through the user's tweets. The strength of negative emotion conveyed in the tweet is heightened with the increased usage of negative words.

This detailed process combines Twitter data mining, sentiment analysis, and machine learning classification to predict and interpret depression-related sentiments in social media posts.

3.2 SOME MORE SNAPSHOTS AND WORKFLOWS

1	weaksbj	1	abandoned	adj	n	negative
2	weaksbj	1	abandonment	noun	n	negative
3	weaksbj	1	abandon	verb	y	negative
4	strongsbj	1	abase	verb	y	negative
5	strongsbj	1	abasement	anypos	y	negative
6	strongsbj	1	abash	verb	y	negative
7	weaksbj	1	abate	verb	y	negative
8	weaksbj	1	abdicate	verb	y	negative
9	strongsbj	1	aberration	adj	n	negative
10	strongsbj	1	aberration	noun	n	negative
11	strongsbj	1	abhor	anypos	y	negative
12	strongsbj	1	abhor	verb	y	negative

Fig 3.2.1 – Dictionary Snapshot

```
1 [{"created_at": "Fri Jun 02 00:04:00 +0000 2017", "id": 870430762255953920, "id_str": "870430762255953920", "text": "Hey, look - I found I"}
2
3
4 {"created_at": "Fri Jun 02 00:04:02 +0000 2017", "id": 870430770141253632, "id_str": "870430770141253632", "text": "RT @shannonpurser: D"}
5
6
7 {"created_at": "Fri Jun 02 00:04:02 +0000 2017", "id": 870430771592413187, "id_str": "870430771592413187", "text": "RT @HRoyalThighness:"}
8
9
10 {"created_at": "Fri Jun 02 00:04:02 +0000 2017", "id": 870430772800479233, "id_str": "870430772800479233", "text": "How to Deal with Str"}
11
12
13 {"created_at": "Fri Jun 02 00:04:03 +0000 2017", "id": 870430776432644096, "id_str": "870430776432644096", "text": "RT @COCONUTOILBAE: w"}
14
15
16 {"created_at": "Fri Jun 02 00:04:04 +0000 2017", "id": 870430779439841280, "id_str": "870430779439841280", "text": "@sabbunny I went the"}
17
18
19 {"created_at": "Fri Jun 02 00:04:04 +0000 2017", "id": 870430780425723904, "id_str": "870430780425723904", "text": "RT @loopzoo: Hello?"}
20
21
22 {"created_at": "Fri Jun 02 00:04:04 +0000 2017", "id": 870430781927288833, "id_str": "870430781927288833", "text": "Gossip Girl and Dona"}
23
24
25 {"created_at": "Fri Jun 02 00:04:05 +0000 2017", "id": 87043078467803104, "id_str": "87043078467803104", "text": "RT @lucasbavid: Me: "}
26
27
28 {"created_at": "Fri Jun 02 00:04:08 +0000 2017", "id": 870430796485726208, "id_str": "870430796485726208", "text": "I'm glad I understand"}
29
30
```

Fig 3.2.2 – Tweet Data Snapshot

ID	Sentiment
870430762	-1
870430770	-1
870430771	-1
870430772	-1
870430776	0
870430776	-1
870430776	-1
870430780	-1
870430781	0
870430784	0
870430796	0
870430796	0
870430796	-1
870430800	0
870430804	0
870430804	0
870430807	-1
870430807	1
870430814	1
870430815	1
870430815	0
870430820	0
870430822	0
870430825	-1
870430838	0
870430838	1
870430840	-1
870430841	1
870430845	-1

Fig 3.2.3 – Output Snapshot

3.3 RESULTS FROM DIFFERENT CLASSIFIERS MORE ELABORATED

❖ Naive Bayes:

Interpretation:

Naive Bayes achieved a relatively high accuracy of 93.79% on the dataset.

The model's completion time was very fast, taking only 0.59779 seconds to process.

❖ Decision Tree:

Interpretation:

Decision Tree performed exceptionally well with an accuracy of 98.56%.

The model took a bit more time compared to Naive Bayes, with a completion time of 3.40457 seconds.

❖ Support Vector Machine (SVM):

Interpretation:

SVM achieved an accuracy of 50.0%, which suggests it performed no better than random chance on this particular dataset.

The completion time was relatively high at 29.83311 seconds, indicating SVM might be computationally expensive for this dataset.

❖ K Neighbors:

Interpretation:

K Neighbors achieved a good accuracy of 81.46%.

The completion time was moderate, taking 7.99048 seconds.

❖ Random Forest:

Interpretation:

Random Forest performed poorly with an accuracy of 49.10%.

Similar to Naive Bayes, the completion time was fast, taking only 0.60994 seconds.

Overall Summary:

Decision Tree stands out with the highest accuracy among the models.

Naive Bayes and Random Forest completed quickly, while SVM took the longest.

SVM's low accuracy suggests it might not be well-suited for this specific dataset.

3.4 SAMPLE TWEETS – MORE ELABORATED

❖ ****Tweet Information:****

- Text
- Created At
- Tweet ID
- Source
- Truncated

❖ ****User Information:****

- User ID
- User Handle
- User Name
- User Location
- User Description
- User Follower/Friends/Statuses Count
- User Profile Image URL
- User Background Image URL

❖ ****Tweet Entities:****

- Hashtags/URLs/User Mentions/Symbols

❖ ****Tweet Engagement:****

- Retweet Count
- Favorite Count

❖ ****Other Metadata:****

- Filter Level
- Language
- Timestamp (ms)

❖ ****User Profile Styling:****

- Profile Styling (Colors, etc.)

This detailed breakdown provides insights into the content, user, and engagement aspects of the given tweet. It's a snapshot of a user's expression related to social anxiety, capturing the context and metadata associated with the tweet. The additional user information gives a glimpse into the user's profile and online presence on Twitter.

```
=====

Retrieving TXT File
Retrieving Successfull

Recovering Data Teets
Data Tweets Recovered

Reading Dictionary
Dictionary Preparation Done

Processing please wait...
```

Fig 3.4.1 – Processing

```
Processing time: 109.67550182 Seconds

=====
Processing Finish
=====
Data Saved!
=====

In [23]:
```

Fig 3.4.2 – Preprocessing Done Figure

```
stream.filter(track=['depression', 'anxiety', 'mental health', 'suicide', 'stress', 'sad'])
```

Fig 3.4.3 – Sample Tweet

❖ **Accuracy:**

- **Decision Tree:** 98.55658748040587%
- **Naive Bayes Accuracy:** 93.79416648429645%
- **K Neighbors:** 81.464122923447%
- **Random Forest:** 49.1037137743686%
- **Support Vector Machine:** 50.0001%

❖ Completion Time:

- **Decision Tree:** 3.40467 Seconds
- **Random Forest:** 0.60984 Seconds
- **Naive Bayes Accuracy:** 0.59788 Seconds
- **K Neighbors:** 7.99038 Seconds
- **Support Vector Machine:** 29.83321 Seconds

```
PS F:\BTP\clone1\Classification-of-Depression-on-Social-Media-Using-Text-Mining> & C:/Users/palaa/AppData/Local/Programs/Python/Python312/python.exe f:/BTP/clone1/Classification-of-Depression-on-Social-Media-Using-Text-Mining/Accuracy_checker.py

Naive Bayes Accuracy :
93.79406648429645 %
Completion Speed 0.26741

Decision tree Accuracy :
98.55668748040587 %
Completion Speed 3.57236

Support vector machine Accuracy :
93.62738823407057 %
Completion Speed 12.45439

Kneighborsclassifier Accuracy :
81.464022923447 %
Completion Speed 2.13076

Random Forest Accuracy :
46.92481190883222 %
Completion Speed 0.48772
```

Fig 3.4.4 – Accuracy Check And Completing Time

```
PS F:\BTP\clone1\Classification-of-Depression-on-Social-Media-Using-Text-Mining> & C:/Users/palaa/AppData/Local/Programs/Python/Python312/python.exe f:/BTP/clone1/Classification-of-Depression-on-Social-Media-Using-Text-Mining/InsertTweetDemo.py

Input your tweet :
i am sad.

*****
Negative
*****
```

Fig 3.4.5 – Negative Tweet Demo

```
PS F:\BTP\clone1\Classification-of-Depression-on-Social-Media-Using-Text-Mining> & C:/Users/palaa/AppData/Local/Programs/Python/Python312/python.exe f:/BTP/clone1/Classification-of-Depression-on-Social-Media-Using-Text-Mining/InsertTweetDemo.py

Input your tweet :
i am so happy today.

*****
Neutral
*****
```

Fig 3.4.6 – Neutral Tweet Demo

```
PS F:\BTP\clone1\Classification-of-Depression-on-Social-Media-Using-Text-Mining> & C:/Users/palaa/AppData/Local/Programs/Python/Python312/python.exe f:/BTP/clone1/Classification-of-Depression-on-Social-Media-Using-Text-Mining/InsertTweetDemo.py

Input your tweet :
this is good in every way

*****
Positive
*****
```

Fig 3.4.7 – Positive Tweet Demo

CHAPTER 4

CONCLUSION AND SCOPE FOR FUTURE WORK

4.1 Task, Achievement, and Possible Beneficiaries:

Task:

The "Classification of Depression Using Text Mining" project aims to tackle the prevalent issue of mental health by leveraging advanced technologies. The primary task revolves around developing a system that can effectively classify and predict early signs of depression through text mining on social media platforms like Twitter and Facebook. The project involves intricate steps, including data collection, preprocessing, model training, and sentiment analysis. The main objective is to utilize machine learning, specifically text mining, to analyze large amounts of data generated on social media and produce meaningful outcomes related to mental health.

Achievements:

The achievement of developing the Depression Classification project is multifaceted and holds significant implications for mental health awareness and support.

❖ **Next-Gen Mental Health Analysis:**

- The project represents a pioneering step in the development of next-generation mental health applications. By employing advanced text mining techniques, the system supports the classification of depression, bringing a technological solution to a critical societal issue.

❖ **Multifunctional Application:**

- The project incorporates machine learning methodologies such as Naive Bayes, Decision Tree, Support Vector Machine, K-neighbors, and Random Forest to achieve accurate depression classification. The inclusion of diverse classifiers ensures a comprehensive approach to sentiment analysis.

❖ **User-Friendly Interface:**

- The application is designed with a user-friendly interface, ensuring accessibility for users with varying levels of technical expertise. The emphasis on user-friendliness encourages widespread adoption and utilization.

❖ **Potential of Text Mining:**

- The project showcases the potential of text mining in mental health analysis. By analyzing language patterns on social media, the system demonstrates the capability to detect and classify early signs of depression, contributing to timely interventions and support.

Possible Beneficiaries:

❖ General Users:

- Individuals facing mental health challenges can benefit from the Depression Classification project. By providing an early indication of depression through text analysis, users can seek support and intervention.

❖ Healthcare Professionals:

- Mental health professionals can utilize the application as an additional tool for early detection and monitoring of depression among individuals. The system's accuracy and efficiency contribute to a more comprehensive understanding of users' mental well-being.

❖ Social Media Platforms:

- Social media platforms may consider integrating similar systems to contribute to user well-being. The insights generated by the project can enhance platform features related to mental health support and intervention.

❖ Researchers and Academia:

- The project opens avenues for further research in the intersection of technology and mental health. Academics and researchers can explore the methodologies employed, contributing to advancements in the field of mental health analysis.

In summary, the Depression Classification project not only presents a technological solution for mental health awareness but also emphasizes user-friendliness and potential societal impact. The beneficiaries range from individuals seeking mental health support to professionals in the healthcare sector and researchers exploring innovative approaches to mental health analysis.

4.2 Review of Contributions:

The "Classification of Depression Using Text Mining" project presents valuable contributions in the domain of mental health analysis and sentiment classification through text mining techniques. The project's contributions can be highlighted as follows:

❖ Advanced Text Mining Techniques:

- The project employs sophisticated text mining methodologies to analyze data from social media platforms, particularly Twitter and Facebook. By utilizing algorithms such as Naive Bayes, Decision Tree, Support Vector Machine, K-neighbors, and Random Forest, the system achieves accurate sentiment analysis, contributing to the advancement of text mining applications in mental health classification.
- ❖ **User-Friendly Interface:**
 - Similar to the emphasis on a user-friendly interface in other projects, this system prioritizes accessibility for users. The intuitive design ensures that individuals with varying levels of technical expertise can easily interact with the system, promoting broader adoption and engagement.
- ❖ **Multifunctional Application:**
 - The integration of multiple machine learning classifiers, including Naive Bayes, Decision Tree, Support Vector Machine, K-neighbors, and Random Forest, adds versatility to the application. This multifunctionality enhances the robustness and effectiveness of the system in sentiment analysis and depression classification.
- ❖ **Potential for Early Detection:**
 - The project's focus on identifying early signs of depression through text analysis contributes significantly to proactive mental health interventions. Early detection is crucial for timely support, and the project addresses this need by leveraging social media posts for sentiment analysis.
- ❖ **Open-Source Availability:**
 - Similar to the open-source model in other projects, this system embraces openness by providing free access for usage, modification, and distribution. The open-source nature encourages collaboration, innovation, and customization, allowing users and organizations to adapt the software to their specific requirements.
- ❖ **Societal Impact:**
 - The overall contributions of the project have the potential to make a positive impact on how individuals and organizations approach mental health awareness. By utilizing text mining and machine learning, the project provides a technological solution aligned with the increasing importance of mental health in the digital age.

In summary, the "Classification of Depression Using Text Mining" project contributes significantly to mental health awareness through the application of advanced text mining techniques, user-friendly design, multifunctional classifiers, early detection focus, and an open-source approach. These contributions collectively enhance the project's potential to positively influence mental well-being at both individual and societal levels.

4.3 Scope for future work

The current iteration of the "Classification of Depression Using Text Mining" project lays the foundation for ongoing improvements and advancements. The following future plans have been identified to enhance the system's capabilities and accuracy:

- ❖ **Contextual Semantic Segmentation:**

- The inclusion of contextual semantic segmentation aims to improve the project's understanding of the nuanced meanings within text data. This involves a more in-depth analysis of the context in which words or phrases are used, contributing to a more refined sentiment analysis and depression classification.

- ❖ **Utilization of Stopwords for Increased Accuracy:**

- Integrating stopwords into the model represents a strategic enhancement. Stopwords, commonly used words like "and," "the," or "is," are often filtered out in standard text analysis. However, their inclusion, when contextually relevant, can provide valuable insights into the sentiment of a sentence. Incorporating stopwords can contribute to increased accuracy in sentiment classification.

- ❖ **Integration of Complex Features: N-grams and Part-of-Speech Tags:**

- The introduction of complex features, such as n-grams and part-of-speech tags, aims to enrich the project's understanding of language structures. N-grams involve analyzing sequences of adjacent words, providing insights into the contextual relationships between words. Additionally, part-of-speech tagging enhances the analysis by identifying the grammatical category of each word. Incorporating these features can contribute to a more sophisticated and context-aware sentiment analysis.

Our commitment to refining the project involves strategic integration of semantic segmentation, addressing stopwords, optimizing feature selection, and incorporating complex linguistic features. Aligned with the project's ongoing evolution for enhanced accuracy in classifying depression through text mining, our iterative approach keeps pace with advancements in text analysis and mental health research.

It's crucial to note that this overview simplifies the sentiment analysis pipeline. Real-world projects may involve additional complexities, such as managing imbalanced datasets, multi-class sentiment analysis, and fine-tuning models for specific domains or languages. Techniques like cross-validation and hyperparameter tuning can further enhance model performance.

REFERENCES

1. Nadeem, M., Horn, M., Coppersmith, G., Sen, S. (Year). "Identifying Depression on Twitter." Advanced Placement Research, Jenks High School, Jenks, OK 74037, USA. Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA. Department of Computer Science, University of Tulsa, Tulsa, OK 74104, USA.
2. Ma, L., Whang, Z., Zhang, Y. (May 2017). "Extracting Depression Symptoms from Social Networks and Web Blogs via Text Mining." Lecture Notes in Computer Science. International Symposium on Bioinformatics Research and Applications. DOI: 10.1007/978-3-319-59575-7_29.
3. Violet. (2017). "Classifying Tweets with Keras and TensorFlow." [Publication Date: 9.2.2017]. Retrieved from: [<https://vgpena.github.io/classifying-tweets-with-keras-and-tensorflow/>].
4. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). "Detecting Depression and Mental Illness on Social Media: An Integrative Review." *Current Opinion in Behavioral Sciences*, 18, 43–49. DOI: 10.1016/j.cobeha.2017.07.016.
5. Reece, A. G., & Danforth, C. M. (2017). "Instagram Photos Reveal Predictive Markers of Depression." *EPJ Data Science*, 6(1), 15. DOI: 10.1140/epjds/s13688-017-0128-6.
6. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). "CLPsych 2015 Shared Task: Depression and PTSD on Twitter." *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, 51–60. DOI: 10.3115/v1/W15-24.