SOCIAL MEDIA PROFILE ANALYSIS FOR PREDICTING THE PSYCHOLOGICAL STATE

Thesis

Submitted In Partial Fulfillment Of The Requirement For The Degree of

MASTER OF TECHNOLOGY

IN

COMPUTER TECHNOLOGY

BY

UMESH PAL

University Roll Number: 002010504032

Examination Roll Number: M6TCT23006

Registration Number: 154197 of 2020-2021

Under The Guidance Of

PROF. DIGANTA SAHA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

JADAVPUR UNIVERSITY, KOLKATA

132, Raja Subodh Chandra Mallick Road,

Jadavpur, Kolkata, West Bengal, 700032

JUNE, 2023

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

This is to certify that the dissertation titled SOCIAL MEDIA PROFILE ANALYSIS FOR PREDICTING THE PSYCHOLOGICAL STATE was completed by

Umesh Pal, University RollNo: 002010504032, Examination Roll Number: M6TCT23006, University Registration No: 154197 of 2020-21, under the guidance and supervision of Prof. Diganta Saha, Department of Computer Science and Engineering, Jadavpur University. The findings of the research detailed in the thesis have not been incorporated into any other work submitted for the purpose of earning a degree at any other academic institution.

Prof. Diganta Saha
Department of Computer Science & Engineering
Jadavpur University

COUNTERSIGNED BY

Prof. Nandini Mukherjee

Head of the Department
Department of Computer Science and Engineering
Jadavpur University

COUNTERSIGNED BY

Prof. Ardhendu Ghoshal
Dean, FET

Faculty of Engineering and Technology

Jadavpur University

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled SOCIAL MEDIA PROFILE ANALYSIS FOR PREDICTING THE PSYCHOLOGICAL STATE is a bonafide record of work carried out by UMESH PAL in partial fulfilment of the requirements for the award of the degree Master of Technology in Department of Computer Science and Engineering, Jadavpur University during the period of June 2022 to May 2023 (5^{th} & 6^{th} Semester). It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn there in but approve the thesis only for the purpose for which it has been submitted.

_____

Signature of Examiner

Date:

_____

Signature of Supervisor

Date:

DECLARATION

I certify that,

(a) The work SOCIAL MEDIA PROFILE ANALYSIS FOR PREDICTING THE PSYCHOLOGICAL STATE contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I havetaken permission from the copyright owners of the sources, whenever necessary.

Umesh Pal
Master of Technology
Roll No.: 002010504032
Registration No.: 154197 of 2020-21
Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

ACKNOWLEDGEMENT

First and foremost, I want to express my gratitude to God Almighty for providing me with the strength, wisdom, and capability to go on this amazing adventure and to continue and successfully finish the embodied research work. I'd like to thank Professor Diganta Saha of Department of Computer Science and Engineering at Jadavpur University for his excellent assistance, consistent support, and inspiration during the course of my dissertation. I owe Jadavpur University a great debt of gratitude for providing me with the chance and facilities to complete our thesis.

I am grateful to every one of the teaching and non-teaching personnel whose assistance has made our trip during my research time much easier. I would like to thank my friends for providing me with regular encouragement and mental support throughout our effort.

Last but not the least, my family deserves great recognition. There are no words to express my gratitude to my mother and father for all of the sacrifices you've made on my behalf. Your prayers for me have kept me going thus far.

Umesh Pal
Master of Technology
Roll No.: 002010504032
Registration No.: 154197 of 2020-21
Department of Computer Science and Engineering
Jadavpur University, Kolkata

ABSTRACT

In recent years, social media platforms have become a rich source of user-generated content, providing valuable insights into individuals' thoughts, emotions, and mental well-being. SOCIAL MEDIA PROFILE ANALYSIS FOR PREDICTING THE PSYCHOLOGICAL STATE presents a comprehensive review of social media profile analysis techniques used for detecting mental states. The objective is to explore the potential of analyzing social media profiles to identify indicators of various mental health conditions, including depression, anxiety, and stress.

The review encompasses a wide range of methodologies employed in the field, including natural language processing, sentiment analysis, machine learning, and data mining. By leveraging these techniques, researchers have developed approaches to automatically extract valuable information from social media profiles, such as textual posts, images, and social connections, to infer the mental states of individuals.

SOCIAL MEDIA PROFILE ANALYSIS FOR PREDICTING THE PSYCHOLOGICAL STATE highlights the challenges and limitations associated with social media profile analysis for mental state detection, including issues of data privacy, ethical considerations, and the need for context-aware interpretation. Additionally, it discusses the potential benefits and applications of these techniques in various domains, such as early intervention, mental health monitoring, and public health policy.

The main objective of SOCIAL MEDIA PROFILE ANALYSIS FOR PREDICTING THE PSYCHOLOGICAL STATE is to predict the psychology of human mind on suicide using various methods. Methods like Bag-of-Words, Google BERT etc. used tofind out the emotional intelligence on the social media data.

Table of Content

List of Abbreviations

NLP	Natural Language Processing
SA	Sentiment Analysis
AI	Artificial Intelligence
ML	Machine Learning
POS	Part-of-Speech
TA	Text Analysis
NLTK	Natural Language Toolkit
BOW	Bag-Of-Words
TF-IDF	Term Frequency - Inverse Document Frequency
DVT	Document Vector Table
IDF	Inverse Document Frequency

Chapter 1

INTRODUCTION

1.1 Overview

Psychological recognition is based on social media sentiment analysis where data consists of people's views and perspectives from various social media platforms. Twitter is very popular social media platform and India is holding the 3rd position on the number of users since 2020's. Sentiment analysis also referred as opinion mining which deals with extraction of subjective information from text and the text or data is free from the different barriers i.e. age, culture, gender etc. Basically, it is based on wide spectrum of moods. Our ultimate goal is to find out the Psychological recognition and to help people with certain predictions.

1.2 Social Media

Social media refers to online platforms and websites that enable individuals, groups, and organizations to create, share, and interact with content. It is a virtual space where users can connect with others, express themselves, and engage in various activities. Social media platforms allow users to create personal profiles, share text posts, images, videos, and other media, and interact with others through comments, likes, shares, and direct messaging.
The concept of social media revolves around the idea of social networking and communication. It provides a digital space for people to connect with friends, family, colleagues, and even strangers from around the world. Social media platforms facilitate the formation of online communities, interest groups, and virtual networks where users with similar interests or affiliations can interact and share information.

1.3 Application of Social Media Profiles And its features

Social media platforms provide users with the ability to create profiles that represent their online identities. These profiles typically include personal information, interests, and activities. Here are some common elements and features found in social media profiles:

1. **Profile Picture:** Users can upload a profile picture that represents them visually. This image is often displayed alongside the user's username and serves as a visual identifier.
2. **Bio or About Section**: Users can provide a brief description or bio that highlightstheir interests, passions, profession, or any other information they choose to share. It gives viewers a snapshot of who they are.
3. **Personal Information:** Social media platforms may offer sections where users caninput personal details such as name, age, location, education, occupation, relationship status, and more. The extent of information shared depends on individual privacy preferences.
4. **Interests and Hobbies:** Users can indicate their interests, hobbies, and favorite activities. This information helps others with similar interests find and connect with them.
5. **Photos and Albums:** Users can upload and organize photos into albums on their profiles. This allows them to showcase their experiences, events, achievements, or simply share moments from their lives.
6. **Timeline or Feed:** The timeline or feed on a social media profile displays the user's posts, updates, and shared content in chronological order. It provides a comprehensive view of their activities and enables others to engage with their posts.
7. **Friends or Followers:** Social media profiles often display the user's friend or follower count. This indicates the number of connections they have made on the platform. The friends or followers list may be public or limited to specific audiences based on privacy settings.
8. **Social Interactions:** Profiles may show the user's recent interactions, such as comments, likes, shares, or mentions they have received from others. These interactions contribute to social validation and engagement.
9. **Privacy Settings:** Social media platforms allow users to control the privacy settingsof their profiles. Users can decide who can view their profile, post comments, send friend requests, or access their personal information.

1.4 Challenges

Performing profile analysis to predict the psychological state of individuals through social media platforms presents several challenges. Here are some of the key challenges involved:

Limited Profile Information: Social media platforms may not provide comprehensive profile information or may have privacy settings that restrict access to certain details. This limitation can hinder the accuracy of psychological state predictions as the analysis heavily relies on available profile data.

User Misrepresentation: Individuals may intentionally misrepresent themselves on social media, presenting a false image or providing misleading information in their profiles. This deliberate misrepresentation can lead to inaccurate predictions of their psychological state based on the available profile data.

Contextual Understanding: Profile analysis alone may not provide a complete understanding of an individual's psychological state. Contextual factors, such as the timing, situation, or specific content of social media posts, are crucial for accurate predictions. The lack of contextual information can limit the effectiveness of profile analysis for predicting psychological states.

Emotional Masking: Some individuals may choose not to express their true psychological state openly on social media platforms. They may project a positive or neutral image while experiencing negative emotions internally. This emotional masking makes it challenging to accurately assess their psychological state solely based on profile analysis.

1.5 Thesis Objective and Organization

This thesis aims to explore the object detection and tracking by contributing originally to three components:

(a) User Sentiment Analysis

(b) Estimating User Depression.

(a) User Sentiment Analysis:

User sentiment analysis involves analyzing and understanding the sentiment or emotional tone expressed by users in various forms of text, such as social media posts, reviews, feedback, or customer support interactions. The goal is to automatically determine whether the sentiment is positive, negative, or neutral, and sometimes even to identify specific emotions expressed, such as happiness, anger, or sadness.

Applications of user sentiment analysis are diverse and include social media monitoring, brand reputation management, customer feedback analysis, market research, and personalized recommendation systems. Understanding user sentiment enables organizations to gain insights into customer preferences, tailor their products or services, improve customer satisfaction, and make data-driven decisions.

(b) Estimating User Depression:

Estimating user depression is a challenging task that involves analyzing user- generated data to identify potential signs or indicators of depression. This research area focuses on developing computational models and techniques to assess a user' mental health state, specifically related to depression.

Researchers have explored different approaches to estimate user depression, including machine learning methods, natural language processing (NLP) techniques, and data mining from social media platforms, online forums, or electronic health records. These approaches analyze linguistic patterns, sentiment expressions, semantic features, and other contextual factors to identify depressive symptoms and risk factors.

The aim of estimating user depression is to provide timely support, intervention, or resources to individuals who may be experiencing mental health challenges. It can help in early detection, prevention, and treatment of depression, as well as personalized mental health support systems. However, it is essential to consider ethical considerations, privacy concerns, and the need for appropriate interventions when working with sensitive mental health data.

Chapter 2

LITERATURE SURVEY

Despite the fact that sentiment analysis and opinion mining have become oneof the most essential sources in corporate decision making, there are still a few issues that need to be addressed. The following sections describe similar sentiment analysis studies and challenges.

Social networks are increasingly used to share daily activities; moreover, they are also used to connect with others as a form of social support on health issues. Under this scenario, computational approaches have leveraged the information from social media to study the depression as well as to detect users suffering from depression. For example, some studies use representations based on BoW to make the detection at user and post levels. These kinds of representations allow to easily measure and compare the utility of word n-grams to identify depression in posts. Representations based on topics have also been explored. (Nasukawa et al., 2003)

Sentiment analysis is one of the very active fields of research and has accumulated numerous research activities year over year. A common purpose of these studies is pinpointing aperson's attitude in a speech or written text concerning a particular subject matter (Das et al., 2001). The field of sentiment analysis applied to text collected from social media has taken several directions, and our goal is to refine the relevance of the analysis outcome. Admittedly, this field has many real-world implementations, including the focal point of our research work: Predicting psychological behavior various studies have examined its relationship with Natural Language Processing (NLP) and have provided new insight into depression detection.

Many researchers have focused on the Deep learning approach for depression detection (Morinaga et al., 2002). are interested in imbalanced data, in Intensity Estimation via social media, conducted a comparative study, and used a hybrid deep learning model to predict the impact of COVID-19 on mental health from social media. Others focused on the BERT model to study the impact of coronavirus on social life; and target dependent sentiment classification.

Electra as a transformer Approach, recently, has gained interest to detect depression and for emotion classification. Besides, machine learning techniques could mostly offer successful results for depression detection (Pang et al.,2002) studied three factors: Linguistic style, emotional and temporal process; subsequently, they instructed a model to use the factors above separately and in conjunction. Researchers such as measured the latency of detection on social media. They have

implemented their framework for adequate early depression diagnosis. Moreover, they have used numerous text classifying approaches, and results reveal that higher accuracy is the direct effect of appropriate feature selection and its combination with other features.

(Peter D. et al., 2002) mentioned that Emotion not only refers to a specific emotional state, but also refers to all the feelings related to the body, the mind, the senses, and the spirit, which are expressedthrough language. Therefore, Long Short-Term Memory (LSTM), which determines how to use and update the information in the storage unit by capturing the long-term dependencies between sentences to obtain more permanentinformation.

Methodological support for the government and relevant administrative departments. The Hierarchical Attention Network model proposed by Yang for the sentence classification problem uses the attention mechanism to model sentences, which reduces the gradient disappearance problem generated by RNN when processing sequence data to a certain extent (Archak et al., 2007). The DeepMojo model proposed by Feibo has a good effect in the sentiment analysis task of text and emoji expression. Cho designs an adaptive memory and forget structure for each RNN unit, which can learn long-term and short-term features while reducing therisk of gradient dispersion.

Chapter 3

Psychology & its analysis

3.1 Psychology

Psychology is the scientific study of the human mind and behavior. It explores various aspects of human cognition, emotion, perception, personality, motivation, and social interactions. Psychologists use empirical research and theoretical frameworks to understand and explain how people think, feel, and behave.

Psychology encompasses a wide range of subfields, including clinical psychology, developmental psychology, cognitive psychology, social psychology, and many others. Each subfield focuses on different aspects of human experience and behavior, and psychologists employ different methods and approaches to study their respective areas of interest.

Clinical psychology, for example, involves the assessment and treatment of mental disorders and psychological distress. Developmental psychology studies the changes that occur in individuals over their lifespan, from infancy to old age. Cognitive psychology investigates mental processes such as attention, memory, problem-solving, and decision-making. Social psychology explores how individuals are influenced by and interact with others.

Psychologists conduct research in laboratory settings, as well as in real-world environments, to gather data and test hypotheses. They use a variety of research methods, including surveys, experiments, observations, interviews, and case studies, to investigate psychological phenomena.

3.2 RISK FACTORS

- Depression
- Low self esteem
- Mental illness
- Substance abuse or dependence
- Eating Disorders
- Family history of suicide
- Self-mutilation
- Prior suicide attempt

3.3 Suicidal behavior include:

There is no one factor that causes someone to kill herself/himself Most often there is a complicated – and confusing mix – of current stressors and losses + old psychological wounds (which may be hidden or unrecognized) + genetic or biological factors + a psychiatric illness + alcohol or other drugs + a readily available way of dying Converging all at once “perfect horrific storm”

Mental health conditions: Mental illnesses such as depression, bipolar disorder, schizophrenia, anxiety disorders, and substance abuse disorders are often associated with an increased risk of suicide. These conditions can significantly affect a person's thoughts, emotions, and behaviors, leading to a higher vulnerability to suicide.

Previous suicide attempts: A history of prior suicide attempts is a significant risk factor. Those who have previously attempted suicide are at a higher risk of future attempts or completing suicide.

Psychosocial factors: Various psychosocial factors can contribute to suicidal behavior, including:

Stressful life events: Personal crises, such as the loss of a loved one, relationship problems, financial difficulties, academic or work-related stress, and legal issues, can significantly increase the risk of suicide.

Social isolation: Feeling disconnected from family, friends, or community can be a risk factor for suicide. Lack of social support and a sense of loneliness can contribute to feelings of despair and hopelessness.

Childhood trauma: Experiences of abuse (physical, emotional, or sexual), neglect, or other adverse childhood events can have long-lasting effects on mental health and increase the risk of suicide later in life.

Substance abuse: Alcohol and drug abuse, particularly when combined with other risk factors like mental health conditions, can significantly increase the likelihood of suicide.

Access to lethal means: The availability of firearms, medications, or other means ofself-harm can increase the risk of completing suicide. Restricting access to lethal methods has shown to be an effective suicide prevention strategy.

3.4 Myths

Asking someone about suicide will increase the risk of suicide.
Reality: Asking someone about suicide can actually lower anxiety, open up communication, and provide an opportunity for them to share their feelings. This can help in assessing their risk and connecting them with appropriate support.

Only experts can stop a suicide.

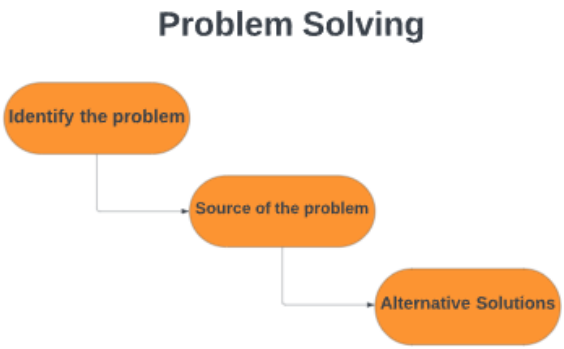
Reality: While mental health professionals play a crucial role in suicide prevention, anyone can make a positive impact. Showing empathy, listening non-judgmentally, and expressing genuine care and concern can be helpful in supporting someone who may be suicidal. Suicidal people don't talk about it.

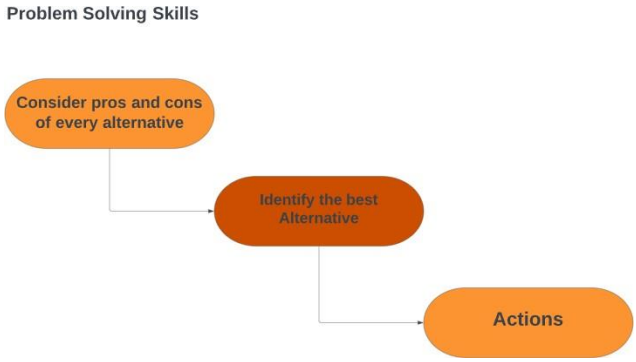
Reality: Many individuals who are contemplating suicide may exhibit warning signs or Give some indication of their intent. These signs can be direct or indirect, such as talking about their thoughts, giving away possessions, or expressing a sense of hopelessness.

Those who talk about suicide don't do it.
Reality: Verbalizing suicidal thoughts should always be taken seriously. It's important not to dismiss or minimize someone's expressed intentions, as they may indeed be at risk. Seeking professional help and offering support is crucial when someone talks about suicide.

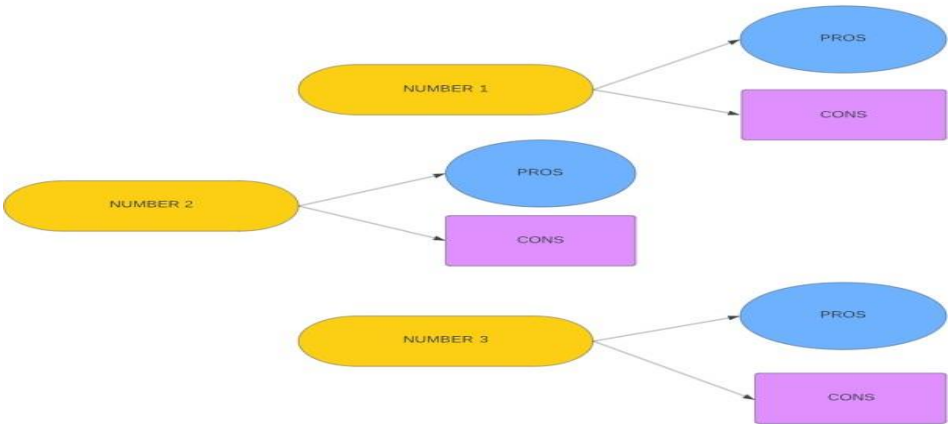
Once a person decides to attempt suicide, no one can change their mind.
Reality: Suicide is often the result of overwhelming emotional pain and despair. Interventions such as professional help, social support, and appropriate treatment can make a significant difference. Many individuals who have survived suicide attempts report that support and intervention saved their lives.

No one can stop suicide.
Reality: Suicide is preventable. When individuals in crisis receive the support and resources they need, the risk of suicide decreases. Interventions can include therapy, medication, crisis hotlines, support groups, and safety planning.





Alternative Solutions



Spend time with family members
Have at least one meal with family members
Involve in travels yearly once (at least)
Network with friends and relationships
Engage in Hobby

3.5 Psycho Linguistics

Psycholinguistics and Sentiment Analysis

Psycholinguistics and sentiment analysis are two distinct fields that can be connected through the study of language and emotions.

Psycholinguistics: Psycholinguistics is the field that examines the psychological and cognitive processes involved in language comprehension, production, and acquisition. It focuses on understanding how humans process and understand language, including the underlying cognitive mechanisms.

Sentiment Analysis: Sentiment analysis, also known as opinion mining, is the processof determining the sentiment or emotion expressed in a piece of text. It involves using natural language processing (NLP) techniques to analyze text and classify it as positive, negative, or neutral. Sentiment analysis can be applied to various domains,such as social media, customer reviews, and feedback analysis.

Relation between Psycho Linguistics and Sentiment Analysis

Language Processing: Both psycholinguistics and sentiment analysis involve studying language processing, albeit from different perspectives. Psycholinguistics focuses on understanding how humans process and comprehend language, while sentiment analysis focuses on automatically analyzing and classifying sentiment in text. Both fields rely on language processing techniques and cognitive models to achieve their goals.

Emotion and Sentiment: Psycholinguistics explores the connection between language and emotion, investigating how emotions are expressed and understood through language. Sentiment analysis, on the other hand, focuses specifically on identifying and categorizing sentiment or emotion in text. The knowledge gained from psycholinguistics can inform the development of sentiment analysis models, helping to enhance their accuracy and understanding of emotional nuances.

Data and Methodology: Psycholinguistics and sentiment analysis often employ different data sources and methodologies. Psycholinguistics may conduct experiments using controlled stimuli and behavioral measures, while sentiment analysis typically relies on large-scale datasets and machine learning algorithms. However, there can be an overlap when sentiment analysis studies incorporate psycholinguistic theories or when psycholinguistic studies incorporate sentiment analysis technique

3.6 Intersection between the two fields

The intersection between psychology and sentiment analysis lies in the field of psycholinguistics. Psycholinguistics is the study of how psychological processes and factors influence language production, comprehension, and communication. Sentiment analysis, on the other hand, involves analyzing and interpreting emotions, attitudes, and opinions expressed in text or speech.

In the context of sentiment analysis, psycholinguistics provides insights into how individuals express and perceive emotions through language. It explores the relationship between linguistic features, such as word choice, sentence structure, and syntactic patterns, and the underlying emotional and cognitive states of individuals.

Psycholinguistic research can help in understanding how different linguistic cues and patterns convey specific emotions and sentiments. For example, certain words or phrases may be associated with positive or negative sentiment, and studying these associations can aid sentiment analysis algorithms in accurately classifying sentiment in text data.

Psychological theories and models related to emotions, such as the appraisal theory or the affective lexicon, can also be applied in sentiment analysis to enhance the understanding and interpretation of emotional content in text. Psycholinguistic factors like context, pragmatics, and individual differences in interpreting sentiment can be considered to develop more nuanced sentiment analysis techniques.
Overall, the integration of psychology and sentiment analysis, specifically within the field of psycholinguistics, allows for a deeper understanding of how language reflects and influences human emotions, attitudes, and opinions, and helps in developing more robust and accurate sentiment analysis techniques.

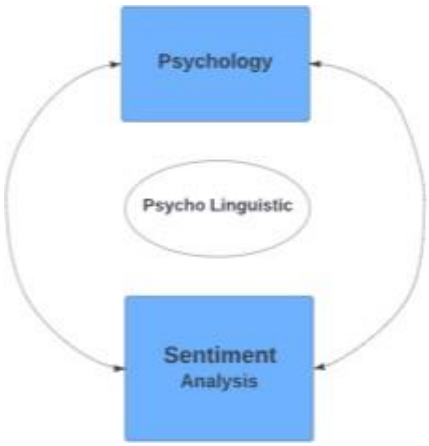


Figure 6.1: Relation of Psychology & Sentiment

Chapter 5

THE PROPOSED WORK

Overview:

The proposed work aims to develop a profile analysis method for predicting an Individual and psychological state. By analyzing various aspects of a person's profile, Including their personality traits, behavioral patterns, and demographic information, the goal is to create a predictive model that can estimate their psychological state Accurately. This work can have applications in fields such as mental health assessment employee well-being evaluation, and personalized interventions.
This chapter of this report addresses the problem statement and outlines the methodology employed for cyberbullying detection using tweets of prominent figures. Cyberbullying on social media platforms, such as Twitter, poses a significant threat to users' well-being and safety. This research aims to develop an effective cyberbullying detection system specifically designed for tweets. We propose a novel approach that combines LSTM models with contextual embedding to capture both the sequential nature of the text and the contextual meaning of words. The proposed system aims to improve the accuracy and robustness of cyberbullying detection on Twitter.

5.1 Algorithm

In this experiment, we have tried out the RNN-based LSTM Neural Network for our datasets. We have tried out the experiment in the English language. Our dataset consists of hate comments (bullying) which are mainly collected from Twitter.

These algorithms are elaborated in six different steps. They are
1. Data Pre-processing
2. Feature Removal
3. Test Train Validation split
4. Describe the model
5. Fit the model
6. Checking the accuracy of the model

1. Data Pre-processing

Input -> .csv file containing raw tweets along with their respective labels.
Output -> .csv file containing cleaned tweets along with their respective labels.
Step 1: Read the tweets of the dataset.
Step-2: Erase the username starting with @.
Step-3: Remove the URL from every single tweet.
Step-4: Swap multiple white spaces with a single white space.
Step-5: The pre-processed dataset is ready.

2. Feature Extraction

Input –> Cleaned Dataset
Output -> Feature extracted and moved for test train split

Step 1: Prepare the maximum sequence length to 250.
Step 2: Describe the embedding dimension to 100.
Step 3: From Keras. Preprocessing we import Tokenizer and apply it to the column of 'Cleaned Tweets'.
Step 4: Now we import texts_to_sequences from keras. Pre-processing and apply it to the column of 'Cleaned Tweets' and save it in a variable.
Step 5: At this time we import the pad sequence from Keras. Preprocessing and applying it over the variable that we get from step 4 and passing the parameter value of padding length that we have initialized in step 1.
Step 6: We can encode the column that is associated with the labels of tweets.
Step 7: Now, Feature extraction is completed.

3. Test Train Validation split

Input –> Feature extracted dataset from part 2
Output–> Feature extracted test. Train and validation of data

Step 1: We have done the test, the train split in a 60:40 ratio over both 'Cleaned Tweets' and 'Label' and saved them in distinct variables.
Step 2: Test. The train and Validation split is now concluded.

4. Defining the model

Input –>None
Output–> A sequential model with an LSTM layer

Step 1: We import Sequential () from Keras and save it as a variable named model.
Step 2: We add an Embedding layer at first with parameters MAX_NB_WORDS, and EMBEDDING_DIM that we have defined in part-2.
Step 3: We have then added a special drop-out layer with a minimum rate = 0.2 to reduce

overfitting.
Step 4: At the moment we add an LSTM layer with 100 memory units with a dropout rate of 0.2 for the input layer, and recurrent dropout rate = 0.2. This is to learn the temporal dependencies among words in the sequences.
Step 5: At present, we add a dense () layer with 'n' output units and the activation function as 'softmax' or 'sigmoid' (according to binary or multiclass classification) which will give the probability distribution over 'n' classes.
Step 6: Now our model is equipped to compile.
Step 7: At this instant, we compile the model with categorical cross-entropy loss or binary cross-entropy, Adam optimizer, and accuracy metric for evaluation during training.

5. Fitting the model

Input –>Pre-processed dataset from part 3 and compiled model from part 4.
Output–>Training of LSTM model using the pre-processed dataset and a trained model.

Step 1: We have fixed our model with the pre-processed dataset and predefined Epoch number and batch size as our choice.
Step 2: Make the validation split 50 % over the test data.
Step 3: Our Model is now trained.

6. Check the accuracy of the model

Input –>Trained model from part 5 and test dataset from part 3.
Output ->Overall Accuracy score of the model, F1 score of each class, and confusion matrix from each class.

Step 1: Using the trained model we predicted the labels of each tweet from a test data set and store them in a variable.
Step 2: From sklearn.metrics we have imported the predefined objects and using the output from step 1 and test data from part 3 we have got our overall accuracy, F1 score, and confusion matrix.
Step 3: Record accuracy, F1 score, and confusion matrix in a Word document.
Step 4: Completed.

5.2 Our proposed work consists of the below steps

Step 1: Prepare suicide ideation scale based on some social media data (Ex.: Twitter) Where ideation scale will contain 8 attributes to predict the suicide tendency.

Step 2: Collect the data from social media (Ex.: Twitter) based on different user Profiles.

Step 3: Apply the data pre-processing which will include the following: Removal of All the links, tags Removal of stop words Removal of numbers.

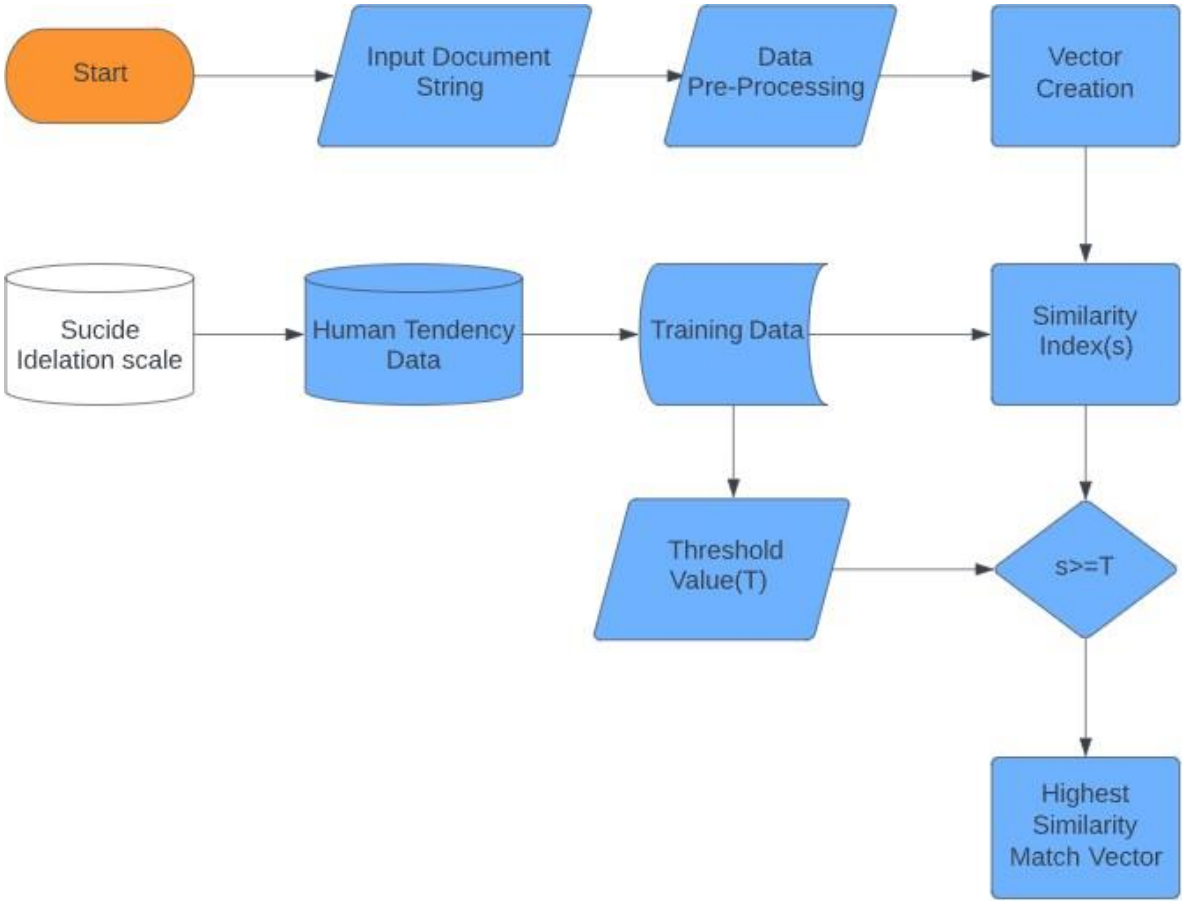
Step 4: Post data pre-processing, will apply different feature extraction methods like BOW (Bag of Words), TF-IDF etc and create the vector out of the text.

Step 5: Calculate the similarity index (s) based on training data

Step 6: find out the vectors where this condition satisfies, s>=T where T is Threshold Value

Step 7: Get the highest similarity vector and assign the corresponding ideation scale to this new vector or the input text.

Proposed Methodology



Chapter 6

Evaluation

6.1 Profile Analysis for Psychological state Report

The aim of this report is to present a profile analysis approach for predicting the psychological state of individuals based on various factors such as textual data, linguistic features, and other relevant attributes. The report explores the utilization of natural language processing (NLP) techniques, specifically the Natural Language Toolkit (NLTK) in Python, to extract meaningful insights from textual data and build predictive models. The report highlights the importance of corpora, preprocessing techniques, sentiment analysis, topic modeling, and machine learning integration in the profile analysis process. The findings of this report contribute to the development of effective methods for understanding and predicting the psychological state of individuals through profile analysis.

BOW: - Bag of Words (BOW) is a commonly used technique in natural language processing (NLP) and information retrieval. It is a simple and effective way of representing text data.
In the Bag of Words model, a document is represented as a collection or "bag" of words, without considering the order or structure of the words in the document. The model assumes that the presence or frequency of words in a document carries important information about the document's content.
The process of creating a Bag of Words representation involves the following steps:
Tokenization: The text is divided into individual words or tokens. Typically, punctuation and other non-alphanumeric characters are removed, and the text is converted to lowercase.
Vocabulary Construction: A vocabulary is created by collecting all unique words from the entire corpus (collection of documents). Each word in the vocabulary is assigned a unique index.

Encoding: Each document in the corpus is encoded as a numerical vector of fixed length, equal to the size of the vocabulary. The vector contains the count or frequency of each word from the vocabulary in the document. Alternatively, binary encoding can be used, where the vector contains a 1 if the word is present in the document, and 0 otherwise.

By representing documents as Bag of Words vectors, it becomes possible to apply various mathematical and statistical techniques for analyzing and comparing textual data. This representation is often used as input to machine learning algorithms fortasks such as text classification, sentiment analysis, topic modeling, and documentretrieval.

Sentence 1: "I love to eat ice cream."
Sentence 2: "I prefer cake over ice cream."

To create a bag of words representation, we first need to tokenize the sentences into individual words. Here are the tokens for the two sentences:

Sentence 1 tokens: ["I", "love", "to", "eat", "ice", "cream"]
Sentence 2 tokens: ["I", "prefer", "cake", "over", "ice", "cream"]

Next, we create a vocabulary, which is a list of all unique words in both sentences:

Vocabulary: ["I", "love", "to", "eat", "ice", "cream", "prefer", "cake", "over"]

Now, we create a vector representation for each sentence based on the vocabulary. The vector will have the same length as the vocabulary, and each element will represent the count of the corresponding word in the sentence.

Sentence 1 vector: [1, 1, 1, 1, 1, 1, 0, 0, 0]
Sentence 2 vector: [1, 0, 0, 0, 1, 1, 1, 1, 1]

In Sentence 1 vector, the first element represents the count of "I" (1 occurrence), the second element represents the count of "love" (1 occurrence), and so on. In Sentence 2 vector, the fifth element represents the count of "ice" (1 occurrence), and so on.

This way, we have transformed the original sentences into numerical vectors, where each element represents the presence or absence of a word in the sentence.
Note that the bag of words representation does not capture the order of the words or any contextual information, but it can be used as a simple and effective way to represent text data for various machine learning tasks, such as text classification or sentiment analysis.

TFIDF:-

TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic used in information retrieval and text mining to evaluate the importance of a term within a document or a collection of documents.

TF (Term Frequency) measures how frequently a term occurs within a document. It is calculated as the number of times a term appears divided by the total number of terms in the document. TF assigns a higher weight to terms that occur more frequently within a document.

IDF (Inverse Document Frequency) measures the rarity or uniqueness of a term across a collection of documents. It is calculated as the logarithm of the total number of documents divided by the number of documents that contain the term. IDF assigns a higher weight to terms that are less common across the collection.

The TF-IDF score for a term in a document is obtained by multiplying its TF value by its IDF value. This score indicates the relevance or importance of a term in a document relative to the entire collection.

TF-IDF is often used in various natural language processing tasks, such as document classification, information retrieval, and text summarization. It helps to identify the key terms that are most relevant to a particular document or a set of documents by giving higher weights to terms that are both frequent within the document and rare across the collection.

We have a collection of three documents: Document A, Document B, and Document C. We want to calculate the TF-IDF score for each term in these documents.

Document A:
"I love dogs."
Document B:
"I love cats."
Document C:
"I love dogs and cats."
Step 1: Calculating Term Frequency (TF)
TF measures how frequently a term appears in a document. We calculate the TF by dividing the number of times a term occurs in a document by the total number of terms in the document.

Document A:
Term frequency for "I": 1/3
Term frequency for "love": 1/3
Term frequency for "dogs": 1/3
Document B:
Term frequency for "I": 1/3
Term frequency for "love": 1/3
Term frequency for "cats": 1/3
Document C:

Term frequency for "I": 1/5
Term frequency for "love": 1/5
Term frequency for "dogs": 1/5
Term frequency for "cats": 1/5

Step 2: Calculating Inverse Document Frequency (IDF)
IDF measures the importance of a term in the entire collection of documents. It is calculated by dividing the total number of documents by the number of documents that contain the term, and then taking the logarithm of that ratio.
Total number of documents: 3
IDF for "I": log(3/3) = log(1) = 0
IDF for "love": log(3/3) = log(1) = 0
IDF for "dogs": log(3/2) ≈ log(1.5) ≈ 0.176

IDF for "cats": log(3/2) ≈ log(1.5) ≈ 0.176
Step 3: Calculating TF-IDF
Finally, we calculate the TF-IDF score for each term by multiplying the TF with the IDF for that term.
Document A:
TF-IDF for "I": (1/3) * 0 = 0
TF-IDF for "love": (1/3) * 0 = 0
TF-IDF for "dogs": (1/3) * 0.176 ≈ 0.059
Document B:
TF-IDF for "I": (1/3) * 0 = 0
TF-IDF for "love": (1/3) * 0 = 0
TF-IDF for "cats": (1/3) * 0.176 ≈ 0.059
Document C:
TF-IDF for "I": (1/5) * 0 = 0
TF-IDF for "love": (1/5) * 0 = 0
TF-IDF for "dogs": (1/5) * 0.176 ≈ 0.035
TF-IDF for "cats": (1/5) * 0.176 ≈ 0.035
In this example, we calculated the TF-IDF scores for the terms "I," "love," "dogs," and "cats" in the three.

Name		Age	Gender	Tell me about your present mood	How are you doing?	Are you in depression?	Do you feel important to your family & friends?	Do you feel lonely?	Have you attended suicide before?	Are you thinking to dosuicide?	Any suicide past history in your family?	Have you written suicidenote in past?
X1		24	Male	Little frustrated and worried. Apart from that I'm usually happy.	Good	No	Yes	Sometimes	No		No	No
X2		26	Male	Perfect	Good	No	Yes	No	No		No	No
X3		25	Male	Happy	Good	No	Yes	No	No		No	No
X4	24	Male	Good	Average	Maybe	Yes	No	No	No		No	No
X5	28	Male	Thoughtful about life and decision making process in life.	Bad	Maybe	Not sure	Yes	No	No		No	No
X6	34	Male	Enjoying	Average	No	Not sure	No	Yes	No		No	No
X7	27	Male	Upset little	Average	Maybe	Yes	Sometimes	No	No		No	No
X8	26	Male	Good	Good	No	Yes	No	No	No		No	No
X9	24	Male	Funny	Average	Maybe	Not sure	Sometimes	No	No		No	No

X10	27	Male	Good	Good	No	Yes	No	No	No	No	No
X11	24	Male	Good	Good	No	Yes	No	No	No	No	No
X12	26	Male	Frustrated	Average	Maybe	Yes	Sometimes	No	No	No	No
X13	38	Female	Good	Good	No	Yes	No	No	No	No	No
X14	23	Male	floccinaucinihilipification	Bad	Yes	No	Yes	No	Yes	No	No
X15	18	Male	Good	Good	No	Yes	No	No	No	No	No
X16	22	Female	Worried about my carrier and health and my family problems	Average	Maybe	Not sure	Sometimes	No	No	Yes	No
X17	24	Male	Bad and fearful	Bad	Maybe	Yes	Yes	No	No	No	No
X18	22+	Male	I am a student in Jadavpur University	Average	No	Yes	No	No	No	No	No
X19	28	Male	Stressful	Good	Maybe	Yes	Yes	No	No	No	No
X20	25	Male	Singing	Good	No	Yes	Sometimes	No	No	No	No
X21	26	Female	Happy	Good	Maybe	Yes	No	No	Yes	No	No
X22	23	Male	Tired	Average	Maybe	Yes	Somet	No	No	No	Yes

							imes				
X23	19	Female	Mixed up	Average	No	Not sure	Yes	Yes	No	No	Yes
X24	23	Male	Anxious	Average	Maybe	No	Yes	No	No	No	No
X25	23	Male	Sad	Average	Yes	No	Yes	No	Yes	No	Yes
X26	25	Male	I am fine.	Good	No	Yes	No	No	No	No	No
X27	26	Female	In a bad mood. Feeling like going somewhere lonely from my regular habitant.	Good	Maybe	Yes	Sometimes	No	No	Yes	No
X28	22	Male	Semi-depressed mood. Research work got stuck in middle and don't know how to progres s anymore. Don't know how to make things right anymore.	Average	Maybe	Yes	Sometimes	No	No	No	Yes
X29	25	Male	Feeling good	Good	No	Yes	Sometimes	No	No	No	No
X30	24	Male	Happy	Good	No	Yes	Sometimes	No	No	No	No
X31	32	Male	Feeling normal with medication	Bad	Yes	Yes	Yes	No	No	No	No
X32	23	Female	Good.	Average	Maybe	Yes	No	Yes	No	No	Yes
X33	27	Male	Good	Good	No	Yes	No	No	No	No	No

X34	26	Male	Don't know it's changing	Good	Maybe	Yes	Sometimes	Maybe	No	No	No
X35	22	Female	Lonely	Good	Maybe	Yes	Sometimes	No	No	No	No
X36	22	Female	Lonely	Good	Maybe	Yes	Sometimes	No	No	No	No
X37	24	Male	Average, not so good not so bad	Average	Maybe	Yes	Sometimes	No	No	No	No
X38	24	Male	Happy	Good	No	Yes	Sometimes	No	No	Yes	No
X39	22	Female	Good	Good	No	Yes	Sometimes	No	No	No	No
X40	24	Male	Confused	Average	No	Yes	Sometimes	No	No	No	No
X41	26	Female	Depressed	Bad	Yes	Yes	Sometimes	No	Maybe	Yes	Yes
X42	18	Male		Good	Yes	Not sure	Sometimes	No	Yes	No	No
X43	29	Male	Happy. And always ready to face and solve whatever problems comesin life.	Average	No	Yes	No	No	No	No	No
X44	21	Female	Happy	Average	No	Yes	Sometimes	No	No	No	No
X45	34	Male	I am soft minded and cool	Good	No	Yes	No	No	No	No	No
X46	25	Male	?	Bad	Yes	Yes	Yes	Maybe	Yes	No	No

Based on actual data to Vector creation and identify the position of 0, 1, 2

	Age	Gender	Sentence	Wish to Live	Wish to Die	Passive suicidal desire	Active suicidal desire	Frequency of Suicide	Control over suicide	Deterrents to active attempt	Suicide note
X1	24	Male	Little frustrated and worried. Apart from that I'm usually happy.	1	0	0	2	0	0	0	0
X2	26	Male	Perfect	1	0	0	0	0	0	0	0
X3	25	Male	Happy	1	0	0	0	0	0	0	0
X4	24	Male	Good	2	2	0	0	0	0	0	0
X5	28	Male	Thoughtful about life and decision making processin life.	0	2	2	1	0	0	0	0
X6	34		Enjoying	2	0	2	0	1	0	0	0
X7	27	Male	Upset little	2	2	0	2	0	0	0	0
X8	26	Male	Good	1	0	0	0	0	0	0	0
X9	24	Male	Funny	2	2	2	2	0	0	0	0
X10	27	Male	Good	1	0	0	0	0	0	0	0
X11	24	Male	Good	1	0	0	0	0	0	0	0
X12	26	Male	Frustrated	2	2	0	2	0	0	0	0
X13	38	Female	Good	1	0	0	0	0	0	0	0
X14	23	Male	floccinaucinihilipilification	0	1	1	1	0	1	0	0
X15	18	Male	Good	1	0	0	0	0	0	0	0
X16	22	Female	Worried about my carrier and health and my family problems	2	2	2	2	0	0	1	0
X17	24	Male	Bad and fearful	0	2	0	1	0	0	0	0
X18	22+	Male	I am a student in Jadavpur University	2	0	0	0	0	0	0	0
X19											
X20	28	Male	Stressful	1	2	0	0	0	0	0	0
X21	25	Male	Singing	1	0	0	2	0	0	0	0
X22	26	Female	Happy	1	2	0	0	0	1	0	0
X23	23	Male	Tired	2	2	0	2	0	0	0	1
X24	19	Female	Mixed up	2	0	2	1	1	0	0	1
X25	23	Male	Anxious	2	2	1	1	0	0	0	0
X26	23	Male	Sad	2	1	1	1	0	1	0	1
X27	25	Male	I am fine.	1	0	0	0	0	0	0	0

X28	26	Female	In a bad mood. Feeling like going somewhere lonely from my regular habitant.	1	2	0	2	0	0	1	0
X29	22	Male	Semi- depressed mood. Research work got stuckin middle and don't knowhow to progress anymore. Don't know how to make things rightanymore.	2	2	0	2	0	0	0	1
X30	25	Male	Feeling good	1	0	0	2	0	0	0	0
X31	24	Male	Happy	1	0	0	2	0	0	0	0
X32	32	Male	Feeling normal with medication	0	1	0	1	0	0	0	0
X33	23	Female	Good.	2	2	0	0	1	0	0	1
X34	27	Male	Good	1	0	0	0	0	0	0	0
X35	26	Male	Don't know it's changing	1	2	0	2	2	0	0	0
X36	22	Female	Lonely	1	2	0	2	0	0	0	0
X37	22	Female	Lonely	1	2	0	2	0	0	0	0
X38	24	Male	Average, not so good not so bad	2	2	0	2	0	0	0	0
X39	24	Male	Happy	1	0	0	2	0	0	1	0
X40	22	Female	Good	1	0	0	2	0	0	0	0
X41	24	Male	Confused	2	0	0	2	0	0	0	0
X42	26	Female	Depressed	0	1	0	2	0	2	1	1
X43	18	Male		1	1	2	2	0	1	0	0
X44	29	Male	Happy. And always ready to face and solvewhatever problems comes in life.	2	0	0	0	0	0	0	0
X45	21	Female	Happy	2	0	0	2	0	0	0	0
X46	34	Male	I am soft minded and cool	1	0	0	0	0	0	0	0
X47	25	Male	?	0	1	0	1	2	1	0	0

Feature wise extraction

Feature	Values								
Wish to live	Bad	0	Frequency of Suicide	No	0				
	Good	1		Yes	1				
	Average	2		Maybe	2	important			
						yes	no	0	
						no	yes	1	
Wish to Die	No	0	Control over suicide	No	0	maybe	not sure	2	
	Yes	1		Yes	1				
	Maybe	2		Maybe	2				
Passive suicidal desire	No	0	Deterrents to active attempt	No	0				
	Yes	1		Yes	1				
	Not sure	2		Maybe	2				
Active suicidal desire	No	0	Suicide note	No	0				
	Yes	1		Yes	1				
	Sometimes	2		Maybe	2				

6.2 Data pre-processing steps:

Data pre-processing shows a vital role in arranging the tweet data for the detection of cyberbullying using LSTM models. Here are the main steps involved in data pre-processing:

I. Text cleaning:

Text cleaning is a vital step in pre-processing of tweet data for cyberbullying detection using LSTM models. Here are several common methods for cleaning text:
Eliminating Special Characters and Punctuation: Tweets often comprise special characters, like emojis, symbols, or emoticons that have not donated to the denotation of the text. These characters can be erased or swapped with proper tokens. Moreover, punctuation marks like commas, periods, and exclamation marks can be detached as they are not crucial for perceiving cyberbullying.
Managing Hashtags and Mentions: In tweets, indications (e.g., @username) and hashtags (i.e., #cyberbullying) are mutual. Depending on the precise necessities, one can select to eliminate them or swap them with any generic term. For instance, one could change references with "USER" and also hashtags with "HASHTAG."
Eliminating URLs: Tweets frequently comprise URLs that can be substituted with a general term like "URL" or wholly detached, as they are contributing to the analysis content.
Managing Numeric Digits: Numeric digits may not be important for spotting cyberbullying in any text. One can pick to eradicate or exchange them with a generic term, like "NUM."
Eliminating Stop words: Stop words are usually the most used words that do not carry ample semantic meaning, like "a,", “and," "the," etc. Eliminating stop words can decrease focus and noise on more expressive words. Though, in specific cases, holding certain stop words may be essential that is liable in this context.
Modifying Spelling and Abbreviations: Text data in tweets frequently comprises misspelled words or acronyms. Applying this method including checking these spelling and intensifying abbreviations can help in standardizing the text and expand the model's understanding of this content.
It has to be noted that the extent of cleaning of all the text and the specific methodologies will vary dependent on the characteristics of the tweet dataset and the requirements of the cyberbullying research project. For analysis of cyberbullying, it is suggested to try diverse text cleaning methods and assess their effects on the performance of LSTM models.

II. Tokenization

Tokenization is a central step in the procedure to analyze text data that include tweets, for detection of cyberbullying using LSTM (Long Short-Term Memory) models. Tokenization denotes the procedure of cutting down any text document into minor units called tokens. In the area of natural language processing (NLP), tokens are characteristically words or subwords. While dealing with these tweets, tokenization develops mainly significantly due to the inadequate count of characters and the unique characteristics of this platform. Here are the steps of tokenization that can be applied in the framework of cyberbullying detection resolutions based on the LSTM models:
Pre-processing: Before tokenization, it is common to perform pre-processing steps such as

removing special characters, hashtags, URLs, and also mentions. These steps are helping to clean the text and eliminate noise that might delay the detection method.
Word-level tokenization: In the tokenization of word-level, the tweet is divided into distinct words or tokens. Every word characterizes a discrete token, and the LSTM model procedures these tokens successively. For instance, the tweet "I dislike you, you are a failure" would be tokenized into ['I', 'dislike', 'you', ',', 'you', "are", 'a, 'failure'].
Tokenization of Sub-word-level: Sub-word-level tokenization is an extra method that splits the text into subwords units that can detect more comprehensive info. This method is mainly beneficial for managing out-of-vocabulary words, slang terms, or abbreviations. Common methods for subwords tokenization comprise Word Piece and Byte Pair Encoding (BPE). For example, the word "unavailable" might be tokenized into ['un', '##ava', '##ail', ##ble'].
Truncation and Padding: Later tokenization, the tweet's tokens are not of equivalent length. For ensuring reliable input dimensions for the LSTM model, truncation or padding is often applicable. Padding includes the addition of special tokens (e.g., <PAD>) to smaller tweets for matching the length of lengthier ones, whereas truncation curtails stretched tweets to a predefined concentrated length.
Representation of Embedding: When tokenization is finished, every token wants to be transformed into a numerical illustration for the LSTM model to generate. This is frequently done by embedding words, that help in mapping words or subwords to impenetrable vectors in an incessant space. Prevalent word embedding methods comprise GloVe, Word2Vec, and Fasttext.
Training and prediction: With the embedded and tokenized data, the LSTM model can be skilled using labeled samples to acquire patterns indicative of cyberbullying. Throughout prediction, different tweets can be tokenized similarly, and the LSTM model can make predictions if they comprise examples of cyberbullying based on the learned outlines.

III. Stemming

Stemming is a text normalization method that is usually used in NLP tasks, containing cyberbullying detection resolutions based on LSTM models. Stemming targets to decrease words to their root or base form, recognized as a stem, by eliminating prefixes and suffixes. The main resolution of stemming is to lessen the dimensionality of the text records and assemble words with similar base meanings, irrespective of their modulations.
Though, in the circumstances of detection of cyberbullying on tweets, stemming might not be as operative or pertinent as it works for NLP tasks. The reasons are-
Unceremonious language and slang: Tweets frequently comprise slang, informal language, formations of creative words, and abbreviations that are the largest on social media stages. Stemming algorithms are intended for standard language and are not handling these variations efficiently. For instance, stemming the word "perfect" to "perfect" might disinvest the anticipated positive sentiment.
Short and context-dependent messages: Tweets are restricted to 280 characters which frequently indicates context-dependent and shortened text. Stemming might not deliver substantial remunerations in a few cases, using the meaning and context of a tweet are often reliant on the whole message, as well as phrases, specific words, and even emoticons.
Loss of data: Stemming can outcome in the loss of info by dropping words to their base forms. In cyberbullying recognition, refined disparities in words or spellings may be significant pointers of insulting or offensive language. By adding stemming, these variations could be

misplaced, possibly decreasing the model's capability to recognize occurrences of cyberbullying precisely.
Although stemming may not be the most active practice for cyberbullying finding on tweets, it's worth noting that these pre-processing stages like eliminating URLs, special characters, and mentions are still appreciated to clean the text documents. Moreover, the use of word embedding can apprehend semantic and appropriate material that can contribute to capturing variations of words and accepting the significance of the text.
Generally, whereas stemming can be convenient in some NLP tasks, its presentation in cyberbullying recognition on tweets based on LSTM models are not delivered important assistance and might even outcome in damaging critical information. It's vital to sensibly deliberate the exclusive characteristics of tweets and the precise necessities of the cyberbullying discovery task at what time to determine pre-processing practices.

IV. Lemmatization

Lemmatization is an additional text normalization method that is used in cyberbullying detection resolutions on tweets based on LSTM models. Lemmatization objects to lessen words to the base or dictionary form, recognized as a lemma, by allowing for the word's context and part of speech. Unlike stemming, which simply eliminates prefixes and suffixes, lemmatization helps in the analysis of words morphologically to certify meaningful alterations. Here are the stages of lemmatization that are applied in the context of cyberbullying recognition on tweets by using LSTM models:
Pre-processing: Earlier lemmatization, it is common to achieve pre-processing stages like eradicating special hashtags, URLs, characters, and mentions. These steps are helping clean the text and eradicate noise that is hindering the detection procedure.
Tagging of Part-of-speech: To achieve lemmatization precisely, the words in the tweet are allocated their consistent part-of-speech (POS) labels. POS tagging classifies every word as an adjective, verb, noun, etc., which is essential for defining the proper lemma.
Lemmatization: Constructed on the POS tags, every word in the tweet is converted into a dictionary or its base form using lemmatization methods. These procedures use language-specific instructions, machine learning algorithms, or dictionaries to achieve the lemmatization procedure exactly.
Truncation and Padding: Afterward the lemmatization, related to the tokenization stage, padding, or truncation are applied to guarantee reliable input extents for the use of the LSTM model.
Embedding demonstration: Once lemmatization is finished, the lemmatized tokens are transformed into numerical representations using word embedding. These representations capture the semantic and contextual material for the words that are critical for the understanding of LSTM models of the tweet.
Training and prediction: With the help of embedded and lemmatized data, the LSTM model is trained using labeled instances to acquire patterns suggestive of cyberbullying. For the period of prediction, new tweets are lemmatized similarly. The LSTM model can create predictions around whether they comprise occurrences of cyberbullying based on the learned configurations.

V. Removal of Step words

The exclusion of stop words is a unique pre-processing stage in numerous tasks of text analysis, comprising cyberbullying finding on tweets based on LSTM models. Stop words are normally used words that are measured to have few semantic meanings and are regularly detached to lessen noise and progress the efficiency and usefulness of algorithms of text processing. Here's how the deletion of stop words is applicable in the framework of cyberbullying detection on tweets consuming LSTM models:

Pre-processing: Earlier eradicating stop words, other pre-processing phases like eliminating hashtags, URLs, special characters, and mentions are normally completed for cleaning the text and eradicating inappropriate information.

Identification of Stop word: A set of stop words that are precise to the particular language is used. These stop words frequently contain prepositions, pronouns, articles, and other mutual words that do not transmit momentous meaning in the analysis.

Elimination of stop words: The stop words recognized in the earlier step are detached from the tweet's text. This can be attained by associating every word in the tweet with all sets of stop words and removing those that can match.

Stuffing and truncation: Afterward the elimination of stop words, like in the preceding phases, truncation or padding are applied to confirm dependable input measurements for the LSTM model.

Embedding demonstration: As soon as stop words are detached, the remaining meaningful tokens are distorted into numerical illustrations using word embedding. These illustrations apprehend the contextual and semantic evidence essential for the LSTM model to comprehend the tweet's content.

Prediction and Training: By the stop words detached and the embedded data are ready, the LSTM model can be qualified using labeled samples to acquire outlines suggestive of cyberbullying. Throughout prediction, new tweets are undertaking similar pre-processing stages, involving the elimination of stop words, and the LSTM model is making predictions grounded on the learned outlines.

The subtraction of stop words are helping to eradicate noise and diminishes the text dimensionality, possibly bettering the productivity and effectiveness of the LSTM model in cyberbullying recognition. By eradicating frequently happening but less instructive words, the model can emphasize more the content-higher terms that might be revealing the cyberbullying performance.

6.2 Result and analysis

Age	Gender	Sentence	Wish to Live	Wish to Die	Passive suicidal desire	Active suicidal desire	Frequency of Suicide	Control over suicide	Deterrents to active attempt	Suicide note
24	Male	littl frustrat worri apart usual happi	1	0	0	2	0	0	0	0
26	Male	perfect	1	0	0	0	0	0	0	0
25	Male	happi	1	0	0	0	0	0	0	0
24	Male	good	2	2	0	0	0	0	0	0
28	Male	thought life decis make process life	0	2	2	1	0	0	0	0
34	Male	enjoy	2	0	2	0	1	0	0	0
27	Male	upset littl	2	2	0	2	0	0	0	0
26	Male	good	1	0	0	0	0	0	0	0

We got the result by using feature extraction and came to know 5 people have negative feedback.

Chapter 7

CONCLUSION & FUTURE WORKS

7.1 Conclusion

The current work on profile analysis for predicting the psychological state has made progress using survey data and a single feature extraction method. While this provides a foundation for understanding users' psychological states, there are several avenues for further improvement and exploration.

Future scopes for enhancing the analysis include collecting data from social media platforms to access real-time user-generated content, incorporating multiple feature extraction methods to capture diverse aspects of users' profiles, and exploring advanced analysis techniques such as natural language processing and machine learning algorithms. Longitudinal analysis can be employed to track changes in users' profiles and psychological states over time, providing a more dynamic understanding of their psychological well-being.

It is essential to validate the predictive models and analysis results by comparing them with external data sources or established psychological assessment tools. This validation ensures the reliability and accuracy of the predictions and allows for evaluating the effectiveness of the developed models.

By expanding the scope of data collection, incorporating various feature extraction methods, and utilizing advanced analysis techniques, researchers can further enhance the accuracy, robustness, and applicability of profile analysis for predicting psychological states. This has the potential to contribute to identifying individuals at risk, providing targeted interventions, and improving mental health support systems.

7.2 Future Works

1. Collect data from social media using APIs: By leveraging APIs provided by social media platforms, researchers can access and collect a larger volume of social media data, including posts, comments, and other relevant information. This will allow for a more comprehensive analysis and a broader representation of users' social media activity.

2. Apply different feature extraction methods: Experimenting with various feature extraction methods can help capture different aspects of the social media data that may be relevant to depression analysis and suicide tendency prediction. This can include not only sentiment analysis but also other linguistic features, topic modeling, network analysis, or even incorporating multimedia elements like images or videos.

3. Better way to find out the similarity index: Developing improved methods for determining the similarity index can enhance the accuracy and effectiveness of identifying individuals with similar patterns or characteristics associated with depression or suicide tendencies. Advanced techniques like text embeddings, graph-based similarity measures, or clustering algorithms can be explored to better identify individuals with similar mental health profiles.

4. Suicide tendency prediction based on the data: Building predictive models to detect suicide tendencies based on social media data is a crucial area of research. This can involve developing machine learning or deep learning models that learn patterns and signals indicative of individuals at risk of suicide. The predictive models can take into account various features such as sentiment, linguistic patterns, network interactions, or other behavioral cues to identify individuals who may require intervention or support.

Bibliography

1. Nasukawa, Tetsuya and Jeonghee Yi. Sentiment analysis: capturing pos-itivity through the use of natural language processing. In Proceedings of the 2nd International Conference on Knowledge Capture, K-CA P-03, 2003.

2.The authors are David, Kushal, Steve Lawrence, and David M. Pennock.Opinion extraction and semantic categorization of product reviews fromthe peanut gallery. In Proceedings of the World Wide Web International Conference (WWW-2003), 2003.

3.Das, Sanjiv and Mike Chen. Yahoo! for Amazon: Mining stock messageforums for market sentiment,2001. In Proceedings of APFA-2001.

4. Morinaga, Satoshi, Kenji Yamanishi, Kenji Tateishi, andToshikazu Fukushima. Web-based product reputation mining 2002. doi:10.1145/775094.775098 Proceedings of the ACM SIGKDD In- ternational Conference on Knowledge Discovery and Data Mining (KDD-2002).

4.Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan have contributed tothis work. Classification of sentiments using machine learning techniques. In Conference on Empirical Methods in Natural Language Processing Pro-
ceedings (EMNLP-2002). 2002.

5.Peter D. Turney's Thumbs up or down? : the application of semanticorientation to the unsupervised classification of reviews. In Annual Meet-
ing Proceedings of the Association for Computational Linguistics (AC L- 2002). 2002.
Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting adjec-tives' semantic orientation. 1997. doi:10.3115/976909.979640 In Proceed-ings of Annual Meeting of the Association for Computational Linguistics
(AC L- 1997).

6.Marti Hearst. Direction-based text interpretation as a refinement of in-formation access, in Text-Based Intelligent Systems, edited by P. Jacobsand published by Lawrence Erlbaum Associates in 1992, pp. 257–274.

7. Identifying subjective characters in story, by Janyce Wiebe 1990. doi:10.3115/997939.998008 In Proceedings of the International Confer-ence on Computational Linguistics (CO LING-1990).

8. .Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis. Show methe money!: deriving the pricing power of product features by miningconsumer reviews. In Proceedings of the ACM SIGKDD Conference onKnowledge Discovery and Data Mining (KDD-2007). 2007.

9.Nasukawa, Tetsuya and Jeonghee Yi. Sentiment analysis: capturing pos-itivity through the use of natural language processing. In Proceedings of the 2nd International Conference on Knowledge Capture, K-CA P-03,
2003.

9.Chen, Yubo, and Xie Jinhong. Online consumer reviews: a new compo-nent of the marketing communication mix. Management Science, 2008.54(3): pp. 477–491. doi:10.1287/mnsc.1070 s.0810

10.Das, Dipanjan. A Survey of Automatic Text Summarization Single-Document Summarization, Language, Volume 4, Issue 4, Pages 1–31,
2007.

11.Dellarocas, C.; Zhang, X. M.; and Awad, N. F. Examining the valueof online product reviews for sales forecasting: The instance of movies 2007 Journal of Interactive Marketing Volume 21 Issue 4 Pages 23-45doi:10.1002/dir.20087.

12. Anindya Ghose and Panagiotis G. Ipeirotis. Designing innovative review rating systems: anticipating the use and influence of reviews. In Proceed- ings of the International Electronic Commerce Conference, 2007.

13. Shivakumar Vaithyanathan, Bo Pang, Lillian Lee, and ShivakumarVaithyanathan. Machine learning algorithms for sentiment classificationget a thumbs up. Conference on Empirical Methods in Natural Language
Processing Proceedings (EMNLP-2002). 2002