ISSN: 2455-2631 March 2023 IJSDR | Volume 8 Issue 3

# Analyzing Mental Health across Twitter Users by Sentiment Analysis

**Abhinav Raj Ansh Agarwal Anuj Kumar Rai Milan Pandey**

Department of Computer Science JSS Academy of Technical Education, Bangalore, Karnataka 560098 INDIA

**Abstract: Social networking services like Twitter provide an abstract representation of one's mental condition. The prevalence of mental health illnesses frequently goes undiagnosed, creating a severe problem that still affects all facets of society. Popular social networking websites can be used to spot recurring psychological trends. These patterns can represent one's daily thoughts and emotions. Our study uses Twitter data to identify people who may be experiencing mental problems and categories them using sentiment analysis techniques based on the language used and certain behavioral characteristics. We present a novel approach for data extraction and concentrate on the study of depression, schizophrenia, anxiety disorders, substance misuse, and seasonal affective disorders in order to address the growing issue of mental disorders. By monitoring users on Twitter for a predetermined amount of time, our technology may be utilized to not only identify but also measure users' advancement. In the long run, this can assist medical practitioners and public health specialists in tracking the signs and patterns of development of mental problems in social media users. The use of social media exacerbates issues with mental health. The consequences of social network use on mental health are summarized in this comprehensive study. Google Scholar databases produced a selection of 50 papers; following the application of various inclusion and exclusion criteria, 16 papers were selected, and each paper was assessed for quality. The remaining eight papers were systematic reviews, three were longitudinal studies, two were qualitative studies, and eight were cross-sectional studies. Anxiety and depression were categorized as two mental health outcomes. Spending time on social media and other related activities has a beneficial impact on the area of mental health. There are, however, a lot of variances because of the cross-sectional design and sampling's technical restrictions.**

**Through qualitative research and vertical cohort studies, the structure of social media effects on mental health has to be better analyzed.**

## I. INTRODUCTION

The World Health Organization (WHO) reports that there are more people suffering from mental diseases than ever before: 322 million individuals worldwide are suffering from depressive symptoms – most of them are women, an increase of 18 percent between 2005 and 2015. These figures encourage effective public health initiatives. For instance, the Mental Health Gap Action Programmed (mhGAP) seeks to increase the number of resources available for people with mental, neurological, and substance use disorders for the countries worldwide. The programmed suggests that even in areas with limited resources, millions of people with depression, schizophrenia, and epilepsy might be treated, stopped from taking their own lives, and start living normal lives with the right care, psychological support, and medication. The epidemic scenario calls for additional study and funding to better comprehend and assist those impacted by such events. All around the world, mental illnesses are now considered a severe problem. According to the Centers for Disease Control and Prevention (CDC), depression will rank as the second greatest global cause of disability by the year 2020. (CDC, 2013). Even substance abuse, violence, and suicide can be caused by mental problems. Most people who suffer from mental disorders frequently go undetected and untreated. Therefore, it is necessary to identify and manage this type of disease. A person's psychology can be studied using social media as a diagnostic technique. Websites for social networking, like Twitter, reflect one's mental state. The Twitter image that follows illustrates a person's emotional state.


**Fig. 1.1 Twitter Image**

Human beings are social creatures. We need one other's companionship to thrive in life, and the calibre of those relationships has a big impact on both our happiness and mental health. Social engagement with others has been demonstrated to shorten lifespan, ease loneliness, ease stress, anxiety, and depression, boost self-esteem, and bring comfort and joy. However, a lack of strong social connections can substantially compromise your mental and emotional health.

Social media platforms like Facebook, Twitter, Snapchat, YouTube, and Instagram are increasingly widely used by individuals throughout the world to connect with one another. While both have benefits, it's important to remember that social media will never completely replace in-person engagement. Only when you are in close physical contact with other people do the hormones that lessen stress and boost happiness, health, and optimism start to flow. Ironically, social media, which is supposed to unite people, can actually increase your sense of loneliness and isolation and aggravate mental health conditions like anxiety and depression if you use it excessively.

It could be time to reevaluate your online behaviours and strike a healthier balance if you're using social media excessively and experience melancholy, dissatisfaction, frustration, or loneliness.

The use of social media has significantly increased in recent years. 72% of Americans in the United States use social media, according to the Pew Research Center.

People utilise social networking platforms to discuss their political opinions, acquire news, and stay in touch with family and friends. Some studies are now considering the long-term consequences of social media use in light of this.

There is no extensive research examining the consequences of social media use because it is still relatively new. However, multiple research shows that social media has a variety of negative effects on mental health.

Many Americans are more likely to experience social media-related anxiety, depression, loneliness, envy, and even physical illness as a result of their increased reliance on and use of social media.

According to the National Institute for Health and Care Excellence (2015), roughly one in five UK adults have depressive disorders, which are thought to impact more than 300 million people globally (World Health Organisation, 2017). While social support is crucial in assisting sufferers in coping, stigma worries can deter people from sharing their symptoms and physically seeking assistance. Because it lessens some of the negative implications of face-to-face disclosure, social media has become a viable venue for discussion and interaction on health topics, particularly stigmatised diseases (Jamison-Powell et al, 2012). In order to provide support, create awareness, and promote improvements, both individuals and organisations in the health- and social-care sectors actively use commercially available platforms like Facebook (authors/reference after review) and Twitter. Twitter was founded in 2006 and is a form of social media referred to as micro-blogging. It is an internet-based service in which users can post brief updates or messages that are aggregated publicly across users. Users can also choose who they want to follow, so following one another is not always reciprocal (Murthy, 2012). Twitter messages are brief; up until 2017, the character limit per message was 140. The hashtag symbol # enables searchable and follow able keyword tagging for tweets (Zappavigna, 2015). When combined, these microblogging qualities enable a variety of communicational tasks, such as enabling people and organizations to use Twitter "as a broadcast medium, marketing channel, diary, social platform, and news source" (Marwick & boyd, 2010: 9).

## II. RELATED WORK

[19]In recent years, many persons with mental health illnesses have received assistance through a combination of in-person interventions and online counseling. Cell phones, social media, and the Internet have revolutionized how people connect, creating new opportunities for providing support. In this new setting, some studies in the literature concentrate on figuring out how well new technologies assist those suffering from mental illnesses. These studies quantitatively analyses the discourse of text-message-based counselling conversations and pinpoint the conversational tactics that lead to the best conservation results.

[20]Recent studies have looked at the issue of spotting depression and suicidal thoughts through social networks. In order to suggest measures for lowering the number of persons afflicted by mental disorders issues, these papers strive to understand how users behave in these social networks.

Numerous studies in the literature have looked into online forums for support and communities for people with mental illnesses on Reddit. People who post on subreddits dealing with schizophrenia, bipolar disorder, and depression have been studied for their writing styles. The analysis demonstrates that users with these illnesses have trouble expressing their thoughts. A strategy for assisting users in the Twitter and Reddit communities is presented via studies conducted by the authors. They suggest ways to determine a user's risk of self-harm, which is associated with users who are depressed.

In order to develop linguistic traits that could be useful for creating further applications to detect persons who require immediate attention, the language used in Reddit communities with a focus on mental health has been examined. In order to create language models that represent social support, which are seen to contain emotional, informational, instrumental, and prescriptive information, users' self-disclosure in Reddit mental illness communities has been characterized. In a similar vein, authors can use the Suicide Watch subreddit's use-case to automatically detect beneficial comments in online posts that appear in suicide watch forums.

We characterize people in mental health-related subreddits using interactions and content analysis of posts and comments in a manner similar to these works. We give a thorough comparison across several subreddits, unlike the studies previously stated. Additionally, we examine the shared information using a more modern and potent topic modelling approach that more clearly identifies the primary subjects discussed by members of these communities.

[16]Different levels of precision in natural language processing tasks have been applied to sentiment analysis. Starting with a classification task at the document level, it was initially tackled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004), then more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009). (Turney, 2002; Pang and Lee, 2004). Because users can post in-the-moment comments and opinions about "anything" on websites like Twitter, microblog data presents new and unique difficulties. Go et al. (2009), (Bermingham and Smeaton, 2010), and Pak and Paroubek are a few of the authors of earlier and more recent studies on sentiment analysis of Twitter data (2010). Go et al. (2009) collect sentiment data via remote learnin. They use tweets as a closing ":)" and ":-)" are regarded as positive and ":(" and ":-(" as negative emotions, respectively, in positive emoticons. Support Vector Machines (SVM), Naive Bayes, and MaxEnt are used to generate the models, and they contend that SVM outperforms other classifiers. Along with parts-of-speech (POS) characteristics in the feature space, they experiment with a Unigram, Bigram model. They note that the unigram model outperforms all other models. POS and Bigrams In positive emoticons, ":)" and ":-") are seen as positive and ":(" and ":-") as negative. Support Vector Machines (SVM), Naive Bayes, and MaxEnt are used to generate the models, and they contend that SVM outperforms other classifiers. Along with parts-of-speech (POS) characteristics in the feature space, they experiment with a Unigram, Bigram model. They highlight the unigram model's superior performance over all others. Bigrams and POS characteristics, in particular, are ineffective.

[17]Similar distant learning paradigms are used for data collection by Pak and Paroubek (2010). However, they carry out a distinct kind of classification task: subjective as opposed to objective. In order to gather subjective data, they follow the same procedure as Go et al. and collect tweets that finish in emoticons (2009). They trawl popular publications' Twitter feeds, such as the "New York Times" and "Washington Posts", for objective data. They report that both POS and bigrams are beneficial (inconsistent with the findings of Go et al. (2009)). But n gram models serve as the basis for each of these techniques. Additionally, they use search queries to gather the incorrect data they require for training and testing sets. Contrarily, we show characteristics that outperform a unigram baseline by a wide margin. We also investigate a different kind of data representation and find that it outperforms unigram models significantly. Another contribution of this study is the fact that our conclusions are based on manually annotated data that is free of all biases that are known to exist. Our data is a randomly chosen sample of tweets that are streaming, as opposed to information obtained through specialized searches. The size of our manually labelled data allows for cross validation. For the purpose of categorizing sentiment in Tweets, Barbosa and Feng have worked significantly harder (2010). They use internet polarity forecasts as noisy labels to train a classifier, then test it on 1000 manually annotated retweets and 1000 manually labelled tweets. They don't, however, mention the method used to gather test results. They suggest combining information like preceding polarity of words and POS of words with syntactic elements of tweets like retweets, hashtags, links, punctuation, and exclamation marks. We build on their strategy by utilizing actual valued previous polarity and fusing it with POS. Our findings demonstrate that the characteristics that link previous orientation of input data with classification model are features that combine words and the components of speech for each. Only slightly, the capabilities of the tweets syntax were helpful.

[18]Gamon (2004) uses the feedback data from the Global Support Services survey to do sentiment analysis. Their study examines the function of linguistic elements like POS tags, among other things. They carry out in-depth feature analysis and feature selection, and they show how the efficiency of the classifier was affected by features of abstract linguistic analysis. In this study, we undertake in-depth feature analysis and demonstrate that the performance of a hard unigram baseline can be achieved using just 100 abstract language features.

In the area of sentiment analysis, numerous researches have been done. A study in this dimension has been undertaken in several studies, with a focus on product reviews in particular [2, 11, 12, 13].

The aforementioned works are primarily concerned with product reviews, which are consistently well-organized sentences that relate to a specific topic. Tweets, which are restricted to 280 characters, have a more informal linguistic style and cover a wide range of topics. Additionally, tweets contain a lot of noise in the form of hashtags, URLs, and emotions. Twitter sentiment analysis is a challenging task because of all these characters.

On classifying Twitter sentiment, researchers [1, 3, 4, 6, 7, 8, 14, 15, 16] are engaged. Good instances of Twitter sentiment can be found in studies [3, 4, 6, 7, 8]. Application based on categorization. In these pieces, the sentiment orientations of online data, such as tweets and reviews, were evaluated to forecast stock market patterns, election results, and box office results.

Go et al[1] .'s method of classifying tweets as favorable and unfavorable employed emotions as noise labels.

They created a publicly accessible corpus of 1.6 million tweets called the Stanford Twitter Sentiment Dataset1. They continued Pang et al. [2]'s classifier training on the basis of this dataset. By combining bigrams and unigrams as features, MaxEnt Classifier gave the best result at 83 percent.

Text pre-processing, especially for tweets, is usually the first step in sentiment analysis [1, 2, 5, 16]. Haddi and others investigated how text pre-processing affected textual data sentiment analysis. To pre-process tweets, Agarwal et al. [16] used certain cutting-edge techniques. The outcome of their study demonstrated how effective text pre-processing techniques can considerably improve the effectiveness of the classifier. There have also been studies like [14], which revealed the original features to improve sentiment classification performance.

Instead of straight integrating the the emotion classifier training incorporates microblogging features, With users, tweets, word unigrams, bigrams, and a few of the microblogging features like hashtags and emoticons as its nodes, Speriosu et al. [14] created a network. The devices are connected solely due to the existence of links between them (for example, users are connected to tweets they created; tweets are linked to word unigrams that they contain; etc.). After that, they employed a label propagation method where sentiment labels were spread throughout the graph from a small selection of nodes that were initially seeded with label information. On the subset of the Twitter sentiment test set, researchers claimed that their label propagation strategy outperformed MaxEnt trained from noisy labels and had an accuracy of 84.7%.[4].

The majority of current research focuses on using three different types of characteristics for sentiment analysis: lexical features, POS features, and microblogging features. The reported results are contradictory. Others [2,6] emphasized the usage of microblogging traits, while some [9,1] advocated the significance of POS tags even without word prior polarity included. The abstract notion that represents each entity in a tweet (such as an iPhone, iPad, or MacBook) will be added as new semantic features proposed in this paper represent a novel class of parameters for sentiment classification (e.g. Apple product). The precision of sentiment analysis is compared to various types of features using unigrams, POS features, and sentiment-topic features. Leveraging such semantic features in the context of sentiment analysis is new, as far as we know.

Social media has undoubtedly evolved over the last ten years. It has positioned itself as a worthwhile topic of attention, and health. There is an expanding body of study on emotions in machine learning.

The domain demonstrates how crucial they are to the research of our online correspondence. The work completed on the job has given us a great deal of insightful information thus far, and provides a solid foundation for the question of how we can achieve much more. Upon reviewing the literature, we propose three specific recommendations, including expanding the use of emotional theories in research, being more precise with the language, and evaluating if Bipolar emotions don't always offer the best explanations; distinct emotions do.

## III. METHODOLOGIES

### Artificial Intelligence

[1]I learned during my study that artificial intelligence is being used to address the current mental health epidemic in America. The truth is that many people find it embarrassing and uneasy talking to a therapist or psychiatrist who is unaware of their vulnerabilities. Teams and businesses are attempting to create AI that can assist and communicate with people who may be struggling with addiction, a mental health issue, or any stressful or triggering event in their lives. Although the idea that artificial intelligence is "replacing" people in mental healthcare may seem strange or unsettling, the outcomes are real. Cognitive behavioral theory is one of the main methods used by mental health practitioners to change a patient's unfavorable thought pattern over the course of multiple counselling sessions. In 2017, clinical psychologists at Stanford University created Woebot, a chatbot that uses cognitive behavioral therapy, a method that has been used by therapists for 40 years to treat sadness and anxiety in patients.

Based on PHQ-9 scores, a typical measure of depression, university students who used Woebot showed close to a 20% improvement in their symptoms in just two weeks. The high degree of participant engagement was one factor in Woebot's effectiveness with the study group. Most people were speaking to the bot virtually every day for a cheap monthly fee of $39, a level of involvement that is simply not possible with in-person counselling. Since this is a relatively new application of artificial intelligence, further research is needed to ascertain just how widely it will be utilized to treat mental health issues.

A number or other type of structured data cannot capture the voice of the people. It's crucial to read, consider, and most importantly comprehend the voice of the people. Brands must be attentive to the various platforms where consumers express their support, disgust, and a wide range of other emotions.

When it comes to the voices of the people when they express their ideas or emotions on social media, AI's natural language technology is particularly helpful since it has the capacity to recognize linguistic nuance. Processing and comprehending the emotions underlying a particular post is becoming more challenging as we advance. People today express themselves in distinctive ways, and the growing use of slang, acronyms, tone, and abbreviations only makes it more challenging to understand the motivation behind a given message. Therefore, it is crucial for AI and its technologies to recognize the specific emotions that a person is conveying.

### Machine Learning

[2]Undoubtedly, a person's emotions, intellect, and ability to communicate with others are all affected by mental illness, which is a health issue. These problems have demonstrated that mental illness has major societal repercussions and necessitates novel prevention and therapeutic measures. Early mental health detection is a crucial step in implementing these techniques. According to Miner et al., medical predictive analytics will fundamentally alter the healthcare industry. The diagnosis of mental illness is

typically based on the patient's self-report, which calls for the use of questionnaires created to identify particular emotional or social interaction patterns. Many people with mental illness or emotional disorders should be able to heal with the right care and therapy.

The rise in mental health issues and the demand for high-quality medical care have prompted research into the use of machine learning in mental health issues. In order to forecast mental health issues, this research presents a current thorough assessment of machine learning algorithms. We must also talk about the difficulties, restrictions, and potential future possibilities for the use of machine learning in the field of mental health.

By utilizing sophisticated statistical and probabilistic methodologies, machine learning tries to create systems that can learn from experience. It is thought to be a very helpful tool for predicting mental health. We were able to create personalized experiences, get crucial information from the data, and create automated intelligent systems. Future events have been predicted and categorized using widely used machine learning methods as the maximum entropy model, artificial neural networks, and naive bayes classifier.

The most frequently used approach in many different sorts of research, studies, and experiments—especially when predicting mental disease in the medical field—is supervised learning in machine learning. All data instances in supervised learning should reflect the names, characteristics, and values. Supervised learning, in more exact terms, is a classification method that makes use of structured training data. Unsupervised learning, however, does not require supervision to make predictions. Handling data without supervision is the primary objective of unsupervised learning. The researchers' ability to use unsupervised learning techniques in the clinical setting is severely constrained.

**Natural Language Processing (NLP)**

[3]Our method uses confusing human language found in social media posts, which makes it extremely difficult to create computers that can accurately determine the intended meaning of text or voice data. The irregularities in human language that take humans years to learn but that programmers must teach natural language-driven applications to recognize and understand accurately from the beginning if those applications are to be useful include homophones, sarcasm, idioms, metaphors, exceptions to the rules of grammar and usage, and changes in sentence structure.

In order to help the computer understand the text and speech data it is absorbing, several NLP activities deconstruct human text and voice data. These are only a few of these jobs: The process of accurately translating voice data into text is known as speech recognition, commonly referred to as speech-to-text. Any programme that responds to voice commands or questions must use speech recognition. The way individuals speak—quickly, slurring words together, with varied emphasis and intonation, in various dialects, and frequently using improper grammar—makes speech recognition particularly difficult. The act of identifying a word's part of speech based on its use and context is known as part of speech tagging, also known as grammatical tagging. In the sentences "I can create a paper plane" and "What make of car do you own?," the word "make" is classified as a verb and a noun, respectively. Word sense disambiguation is the process of choosing one of several possible meanings for a word by using semantic analysis to decide which meaning makes the most sense in the given situation. Word sense disambiguation, for instance, clarifies the difference between the meanings of the verbs "make" and "make the grade" (achieve) and "make a bet" (place). Words or phrases are recognized as useful entities using named entity recognition, or NEM. NEM identifies "Kentucky" as a place or "Fred" as the name of a guy. The task of determining whether and when two words refer to the same item is known as reference resolution. The most typical example is figuring out who or what a certain pronoun refers to (e.g., "she" = "Mary"), but it can also require figuring out a metaphor or idiom that is used in the text (e.g., when "bear" refers to a big, hairy person rather than an animal). Sentiment analysis looks for intangible elements in text, such as attitudes, feelings, sarcasm, bewilderment, and mistrust. Natural language generation is the process of converting structured data into human language; it is frequently referred to as the opposite of voice recognition or speech-to-text. Natural Language Processing, an essential business technology, is needed to glean hidden data insights from social media platforms. Sentiment analysis can be used to analyses language used in social media postings, comments, reviews, and more to gather attitudes and emotions in response to products, promotions, and events. Businesses can use this information to improve their products, advertising strategies, and other aspects of their operations. Raw data is a term used to describe text-based data. This raw data is used for sentiment analysis based on NLP. Many Machine Learning (ML) algorithms, such as SVM (Support Vector Machines), Multinomial Naive Bayes, and MaxEntropy, involve data classification. An essential tool for the backend systems is word embedding. In this approach, words are displayed as vectors. Each word is associated with a single vector, the values of which have been trained to aesthetically and functionally mimic an artificial neural network. Each word vector is then divided into a row of real numbers, where each number stands for a different aspect of the word's meaning. Words with similar semantic meanings or synonyms will have similar or equal vectors.

Text-based data is referred to as raw data. For NLP-based sentiment analysis, this raw data is used. Data classification is used by many Machine Learning (ML) techniques, including SVM (Support Vector Machines), Multinomial Naive Bayes, and MaxEntropy. Word embedding is a crucial tool for the backend systems. Words are shown as vectors in this representation. Each word is connected to a single vector, and the vector values are trained to resemble an artificial neural network both visually and functionally. The next step is to separate each word vector into a row of real numbers, where each number represents a component of the word's meaning. Words that are synonyms or that are semantically comparable will have equal or close vectors.

## IV. ALGORITHMS

**Maximum Entropy**

[4]For natural language applications like information retrieval and speech recognition, Berger et al. (1996) first created the maximum entropy model. This method, which has its roots in information theory, has been effectively used for many years in a variety of domains, including physics and natural language processing. It develops a model that most accurately represents the data currently available, but with the restriction that, in the absence of any new information, the model should maximize entropy. In other words, by maximizing conditional entropy, the model favors a uniform distribution.

When linguistic circumstances are used to forecast linguistic classes, many issues in natural language processing can be seen as linguistic classification issues. To assess the likelihood of a particular linguistic class occurring in a given linguistic environment, maximum entropy models provide a clear way to aggregate various contextual facts.

Models based on maximum entropy (MaxEnt) do not make independent assumptions and are feature-based. With no feature duplication, we can add new features to MaxEnt by using bigrams and phrases. It is possible to estimate any probability distribution using these feature-based modes. Similar to a two-class situation, logistic regression can be used to find a distribution over classes. Belong to the teacher class in a four-way text categorization where we know that 40% of the documents contain the word "teacher." It makes sense that if a document contains the word "teacher," there is a 40% chance that it belongs to the teacher class and a 20% chance that it belongs to one of the other three classes. When the word "teacher" is absent from the document, we can estimate the uniform class distribution at 25% for each.

The aim is to observe some linguistic "context" b ∈ B and anticipate the appropriate linguistic "class" a ∈ A. This is how many natural language processing (NLP) problems can be recast. Building a classier cl: B -> A is necessary to achieve this, and this classier can then be implemented with a conditional probability distribution p, where p(a|b) represents the probability of a "class a" given a "context" b. Contexts in NLP tasks typically consist of at least words, and the precise context varies depending on the task's requirements. For some tasks, context b could be as simple as a single word, while for others, it might be made up of multiple words and their corresponding grammatical labels. Since the words in b are frequently sparse, large text corpora normally contain some information regarding the co-occurrence of a's and b's, but never enough to accurately characterize p(a|b) for all possible (a,b) pairs. The problem then becomes how to accurately estimate the probability model p utilizing the limited evidence regarding the's and b's.

Maximum entropy probability models offer a straightforward way to combine multiple contextual variables in order to evaluate the possibility of a specific linguistic class existing in a specific linguistic environment. Prior to talking, we first show how to characterize the data and include it with a particular kind of probability model inside the maximum likelihood framework, which is followed by an independently motivated interpretation of the probability model within the maximum entropy paradigm.

**Multinomial Naive Bayes Classifier**

[5]In our model, text data classification is an important attribute that has to be dealt with. Text data classification is important because there can be an enormous amount of information available in a single person's social media posts that needs to be analyzed. It is important that we know the context around a certain type of text so that we can perceive the emotions around a person's words and can successfully implement our model to classify the emotions behind his/her posts.

Multinomial Naive Bayes is a popular supervised learning classification that is used for the analysis of categorical text data. The main application of this probabilistic learning technique is in natural language processing. The method, which guesses the tag of a text such as an email or newspaper article, is based on the Bayes theorem. For a given sample, it determines the probabilities of each tag, and then outputs the tag with the highest probability.

An effective algorithm for issues involving numerous classes and text data analysis is naive bayes. Given that the Naive Bayes theorem is built on the Bayes theorem, it is crucial to first comprehend how the latter works.

Based on past knowledge of the circumstances surrounding an event, the Bayes theorem determines the likelihood that the event will occur. It is formulated.

**P(A|B) = P(A) * P(B|A)/P(B)**

Where the probability of class A is calculated when predictor B is already provided.

P(B) = prior probability of B
P(A) = prior probability of A
P(B|A) = occurrence of predictor B given class A probability.

Since the Multinomial Naive Bayes algorithm only calculates probability, it is easy to implement. This method can be applied to both continuous and discontinuous data. The usage of this algorithm is pretty straightforward and is widely used to forecast real time applications plus its scalability and the ease with which it can handle enormous datasets is very important as we are dealing with something very sensitive and a wrong classification can be catastrophic.

For text classification we have used count vectorization and Term Frequency-Inverse Document Frequency (TF-IDF).

**i. Count Vectorization**

[6]Since we know that Count Vectorization involves counting the number of occurrences each word appears in a document, we have used a tool called Count Vectorizer from the Python's Sci-Kit learn library to accomplish this. Example sentence: "**The weather was wonderful today and I went outside to enjoy the beautiful and sunny weather.**" You can tell from the output below that the words "the", "weather", "and "and" appeared twice while other words appeared once. That is what Count Vectorization accomplishes.

It also accomplishes a lot of things. Our model makes use of various Count Vectorizer parameters. One of them is *max_df*. It pretty much shows how many words we want the Count Vectorizer to count. This is pretty useful because we are dealing with a large dataset.

Another parameter is *ngram_range*. It is very useful when we are dealing with a contiguous sequence of words. It helps in acquiring a lot more meaning for the words rather than using the contiguous words as independent features.

**ii. Term Frequency-Inverse Document Frequency (TF-IDF)**
[7]Our methodology also asks us to consider whether a word appears frequently in some texts but less frequently in others, which can be really helpful. Term Frequency-Inverse Document Frequency (TF-IDF) is useful in this situation.

The word "frequency" describes how frequently a phrase, or word, appears in a document.

A term's inverse document frequency indicates how frequently or infrequently it appears in a document.

Inverse document frequency is calculated using the logarithmic function of the number of documents in the set, including any documents in which a term appears more than once. The term frequency is then multiplied by this to obtain a score.

If the TF-IDF score is high, the word is likely to be uncommon and effective at differentiating between papers. Contrary to Count Vectorizer, which merely counts the frequency of a word, this can be quite helpful. The TF-IDF is the technique to utilize if your analysis wishes to look at the discrepancies between documents or the uniqueness of words. Count Vectorizer is the tool to utilize if you're interested in how often a word appears. Once more, it all depends on the subject you wish to study. In order to know which tools to employ, you must comprehend your text. To solve your problem statement, you can absolutely employ a combination of these methods in addition to others.

**Natural Language Toolkit (NLTK)**
[8]For tackling particular NLP tasks, a variety of tools and libraries are available in the Python programming language. The Natural Language Toolkit (NLTK), an open source collection of libraries, tools, and educational resources for developing NLP programmes, contains several of them.

The NLTK offers libraries for many of the above-mentioned NLP tasks as well as libraries for subtasks including sentence parsing, word segmentation, stemming and lemmatization (techniques for removing all but the most essential letters from words), and tokenization (for breaking phrases, sentences, paragraphs and passages into tokens that help the computer better understand the text). Additionally, it has libraries for implementing functions like semantic reasoning, which allows users to draw logical inferences from text-based evidence. Before the machine learning capabilities are even used, NLTK has a variety of functions that may be called with little to no arguments that are helpful in meaningful text analysis. Numerous NLTK tools are useful in getting the data ready for more sophisticated analysis.

**The NLTK corpus and module downloader**: This module defines several interfaces which can be used to download corpora, models, and other data packages that can be used with NLTK.

NLTK can be installed over pip (pip install nltk). After its installation, many components will not be present, and many of the NLTK's features won't be accessible. NLTK's download function is used to select which additional packages need to be installed.

Download () will show an interactive interface that may be used to download and install new packages if it is invoked without any parameters. A graphical interface will be displayed if Tkinter is available; otherwise, a straightforward text interface will be offered. By invoking the download () function and passing it a single argument—the package identification of the particular package to download—individual packages can be downloaded.

To download all packages available: **nltk.download('all')**

To download specific package: **nltk.download('package-name')**

By default, packages are set up in either the current user's home directory or a system-wide directory. However, if another installation target is preferred, it can be specified using the download dir parameter.

**Support Vector Machines (SVM)**
[9]One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name of this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors are the phrase for these extreme circumstances, and as a result, Support Vector Machine is the name of the algorithm.

Any type of vector that encodes any type of data can be subjected to SVM. This means that texts must be converted into vectors in order to take advantage of the capabilities of svm text categorization. Therefore, SVM chooses where to draw the best "line" (or the best hyperplane) that divides the space into two subspaces: one for the vectors that belong to the specified category and one for the vectors that do not; when SVM determines the decision boundary we discussed earlier. Therefore, we will be able to apply the SVM technique to text classification issues and receive extremely good results as long as we can identify vector representations that encode as much information from our texts as possible.

The input and output formats for SVM are predefined. The mental score of an individual (positive/negative) is the output from a vector space as the input.

Text documents should not be used for learning in their native form. They are altered so that they can be entered into machine learning algorithms in the desired format. Preprocessing of text documents is done for this. Next, we do transformation. One dimension will be assigned to each word, and the same dimension will be assigned to identical words. Now a machine learning algorithm is used for learning how to classify documents, i.e. creating a model for input-output mappings.

To build the model we first need to gather the perfect data for training and testing which we have done using **HuggingFace Dataset** (allows us to use posts from a certain individual's twitter account). Then, the data is vectorized using Count Vectorization or Term Frequency-Inverse Document Frequency (TF-IDF). And at last, a linear SVM Model is created which can be used to accurately predict the mental score of a person.

ML algorithms commonly encode examples in vector space, with the majority of the attributes corresponding to words. However, word pairings or a text's word order may include a lot of information, and virtually infinitely many features can be created to improve classification accuracy.

Words in representations are those that appear in texts.Words can, however, have multiple meanings as well as the same meaning in different contexts. The context of a word can help define its appropriate meaning, therefore each word affects the meaning of its surroundings. The typical (computationally practical) representation, however, disregards the word order. The learning and generalization of the input-output mapping is the task of SVM. In the case of text classification, an input set of documents and an output class are both involved.

**Scikit-Learn (Sklearn)**
[10]Python's Scikit-learn (Sklearn) library for machine learning is very practical and reliable. Through a Python consistency interface, it offers a variety of effective tools for statistical modeling and machine learning, including classification, regression, clustering, and dimensionality reduction. This library is based on NumPy, SciPy, and Matplotlib.

We must format the Twitter data before we can use our classifier. Using **sklearn.feature_extraction.text.CountVectorizer,** we will convert the tweets to a matrix or a two-dimensional array of word counts. These vector counts will ultimately be used to train the classifier.

The working data must be imported after we have imported all of the required modules. There is one tweet per line in each text file. We'll split the file into two lists—one for tweets and one for their labels—using the built-in open function. This format was chosen so that we could evaluate how accurate the model we created is. To accomplish this, we test the classifier using unlabeled data rather than labeled data.

Next, we initialize a scikit-learn vector with the CountVectorizer class. Because the data could be in any format, we'll set lowercase to False and exclude common words such as "the" or "and". This vectorizer will transform our data into vectors of features. In this case, we use a CountVector, which means that our features are counts of the words that occur in our dataset. Once the CountVectorizer class is initialized, we fit it onto the data above and convert it to an array for easy usage.

After this, a classifier is used for this dataset and the accuracy of the model is calculated.

**Numerical Python (NumPy)**
[11]The library known as NumPy, or "Numerical Python," contains multidimensional array objects and a selection of procedures for handling those arrays. Arrays can be subjected to logical and mathematical operations using NumPy. The N-dimensional array type known as ndarray is the most significant object defined in NumPy. It describes the collection of items of the same type. A zero-based index can be used to access items in the collection.

Before we can use NumPy, it needs to be imported:

**import numpy as np**

**Pandas**
[12]Working with "relational" or "labeled" data is made simple and intuitive with the help of the Python module Pandas, which offers quick, adaptable, and expressive data structures. It seeks to serve as the essential, high-level building block for using Python for actual, useful data analysis.

Pandas is a versatile tool that works well with a variety of data types, including tabular data with heterogeneously-typed columns, ordered and unordered time series data, arbitrary matrix data with row and column labels, and other sorts of observational or statistical data sets. The bulk of common use cases in finance, statistics, social science, and many branches of engineering are handled by the two main data structures of pandas, Series (1-dimensional) and DataFrame (2-dimensional).

**Python Pickle**
For serializing and de-serializing a Python object structure, use the pickle module in Python. In Python, any object can be pickled in order to save it to disc. Pickle "serializes" the object before sending it to a file, which is what it does. A Python object (list, dict, etc.) can be turned into a character stream by pickingling. This character stream is supposed to contain all the data required to recreate the object in another Python function.

Pickle constructs arbitrary Python objects by invoking arbitrary functions, that's why it is not secure. However, this enables it to serialize almost any Python object that JSON and other serializing methods will not do.

Unpickling an object usually requires no "boilerplates". So, it is very suitable for quick and easy serialization.

**Data Pre-Processing**
While we are trying to analyze the mental health of a person using the person's social media posts, it is often tricky for the machine to understand the language used. Human language is tricky and is unstructured in nature. Finding insights from text data is not as straightforward as structured data and it needs extensive data pre-processing.

Features are not clearly provided in text data, in contrast to structured data. So, in order to extract features from the text data, a method is required. To determine if a word is present or absent in a phrase, one approach is to view each word as a feature and develop a metric. The bag-of-words (BoW) model is used to describe this. Therefore, each sentence is viewed as a collection of words. Each sentence is referred to as a document, and the corpus is the collection of all documents.

**i. Bag-of_words (BoW) Model**

An approach to extracting features from text for use in modeling, such as with machine learning techniques, is known as a bag-of-words model, or BoW for short. The method is really straightforward and adaptable, and it may be applied in a variety of ways to extract features from documents.

By taking into account the frequency of words in the provided document, the Bag of Words model is utilized to represent the text in numbers. We can visualize a phrase as a collection of words, just like the term suggests (a string of numbers).

It considers only two things-

**1. A vocabulary of words**

**2. Presence (or frequency) of a word in a given document** ignoring the order of the words (or grammar).

We divided our model into training and test datasets before using bag-of-words. The bag-of-words representation of our training and testing dataset is then obtained by using **Count Vectorizer** from the sklearn package. As a result, we will train our vectorizer on the training set of data before applying it to the test set.

**ii. Tokenization**

Tokenization is the first step in every NLP pipeline. It has a big impact on the rest of your pipeline. Tokenization is the process of separating natural language text and unstructured data into data units that can be viewed as discrete components. Token occurrences in a document can be used directly as a vector to represent the document. With this, unstructured text or strings are instantaneously transformed into numerical data formats that are suitable for machine learning. They can also be used directly by a computer to launch beneficial responses and actions. Or, they might be used as features in a pipeline for machine learning to start more sophisticated actions or judgments.

Sentences, words, letters, and subwords can all be broken apart using tokenization. Sentence tokenization refers to the process of dividing a text document into sentences. For words, we call it word tokenization.

Along with a collection of text processing modules for categorization, tokenization, stemming, and tagging, NLTK offers user-friendly interfaces for more than 50 corpora and lexical resources, including WordNet.

Words can be easily tokenized using the tokenize module of NLTK.

The relevant functions need to be imported from the NLTK library:

**Import nltk from nltk.tokenize import word_tokenize**

**iii. Stopwords**

Many words that are frequently used in social media posts are unrelated to the data they are used with and don't provide any further depth to the phrase. Stop words are these phrases. The words like 'if', 'but', 'we', 'he', 'she', 'they', etc. do not contribute any meaning and are removed without changing the semantics of a text and doing so often improves the performance of a model.

Stopwords can be imported from the nltk corpora:

**From nltk.corpus import stopwords**

**iv. Normalization**

Correct processing must be applied to words that appear different due to case or are written in a different way but have the same meaning. These terms are given the same treatment thanks to normalization techniques. For instance, changing the case of all the text or transforming numerals to their word equivalents.

The process of transforming a token into its original form is known as normalization. The inflectional form of a word is eliminated during the normalization process in order to produce the base form.

Normalization is useful for removing variations from a text and lowering the amount of unique tokens that are there, also editing the wording by eliminating unnecessary details.

**Lemmatization** and **stemming** are two often used techniques for normalizing.

**Stemming** is a simple rule-based technique which removes inflationary forms from a given token and returns them to their word stem, base, or root form (generally a word written form). For example – laughing, laughed, laughs, laugh all will become laugh after the stemming process.

**Lemmatization**, a systematic process, removes the inflectional form of a token and transforms it into a lemma. Word structure, vocabulary, part of speech tags, and grammar relations are all used during the process of lemmatization.

A root word known as a lemma is the product of lemmatization. For instance, "am," "are," and "is" will be changed to "be." Running, runs, and "ran" will all be changed to "run."

Lemmatization also allows one to specify the desired term's part of speech because it is a systematic process.

Additionally, lemmatization can only be done if the word being used contains the correct parts of speech tag. For instance, the word running will become run if we attempt to lemmatize it as a verb. However, the same word running won't change into a noun if we attempt to lemmatize it.

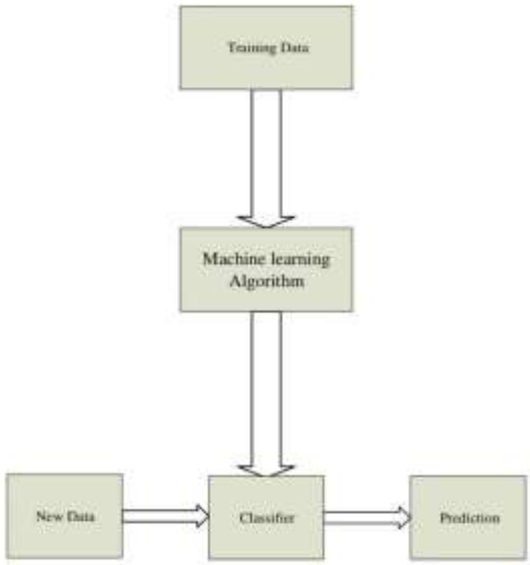


**Fig. 1.2 Flow Diagram 1**

**Regular Expressions (RegEx)**
[13]When we work with text data, it virtually never comes to us in the form we want. There may be words in the text that we want to remove, unnecessary punctuation, HTML or URLs that can be removed, and dates or number entities that can be made simpler.

RegEx is an extremely potent programming tool that may be used for many different things, including the extraction of textual features, replacing strings, and manipulating strings in general.

In order to locate substrings within a given text, one uses a regular expression, which is a collection of characters or a pattern. For instance, obtaining email addresses, phone numbers, or other information from a big unstructured text content. An efficient way to match or replace patterns inside of a string is to utilize a regular expression, which is a text string that represents a search pattern.

In essence, regular expressions make it simple to extract, substitute, and perform a number of other string manipulation operations from any string that contains a pattern. Since almost all widely used programming languages enable using regexes, regular expressions are essentially a language unto themselves. They even have their own compilers.

For implementing regular expressions, we have used Python's **re** package and it can be imported as:

**import re**
ReGex provides a lot of functionalities such as searching patterns in a string (match and findall function of the re package is used), searching alphabets (using the match function and providing the alphabets as the input which we want to find), matching string from the start (carrot key ‘^’ followed by the word to match with the search function is used), matching strings from the end (required word followed by the dollar sign ‘$’), substituting text in a string (sub function with the words that is going to be replaced and the word that is going to replace the word), removing digits from a string (the regex expression to find digits in a string is ‘\d’),

removing alphabet letters from a string, removing word characters (to remove all the word characters (letters and numbers) from a string and keep the remaining characters, we can use the '\w' pattern in your regex and replace it with an empty string of length zero), removing non-word characters (\w pattern can be used to remove all the non-word characters) and a lot more.

Our project deals with removing hashtags, special symbols, emojis, multiple spaces, removing digits from a string and removing loads of unnecessary information.

To achieve this, we have used the following code:

**'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+',""**

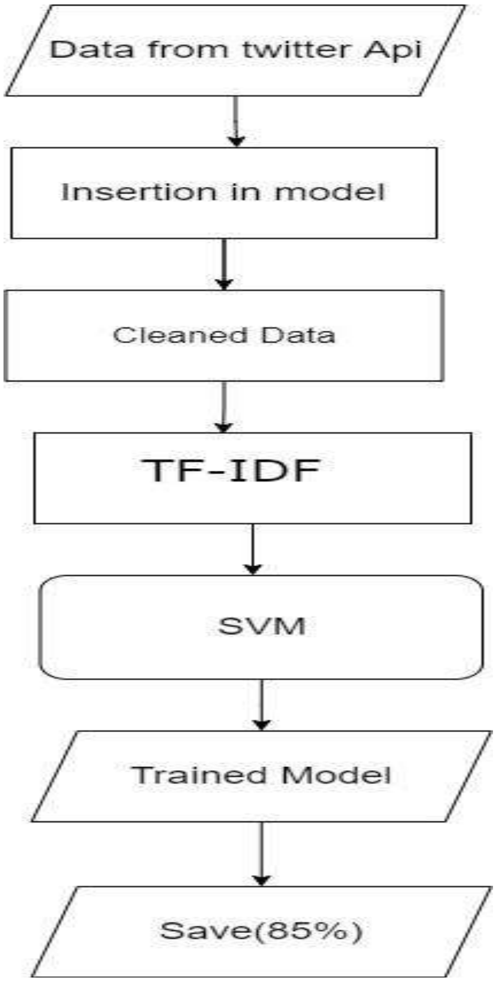


**Fig. 1.3 Flow Diagram 2**

**HuggingFace Dataset**
[14]As our model deals with a large number of social media posts from different users, we needed a dataset that can provide us with ready-to-use datasets for machine learning models with fast, easy-to-use and efficient data manipulation tools and the HuggingFace hub provides us with just that.

To perform text categorization, information retrieval, question-and-answer, translation, text generation, and summarization, HuggingFace offers hundreds of pretrained models. Transformers offer easily downloadable APIs that apply pretrained models on a text to optimize users' datasets.

For our model, we have used the Emotion dataset which is a dataset of english twitter messages with six basic emotions: **anger, fear, joy, love, sadness,** and **surprise**.

The data fields are the same among all splits.

In default case, **Text:** a string feature

**Label:** a classification label, with possible values including anger (0), fear (1), joy (2), love (3), sadness (4), surprise (5).

| text (string) | label (class label) |
|---|---|
| i am now nearly finished the week detox and i feel amazing | 5 (surprise) |

**Fig. 1.4 HuggingFace Dataset Example**

### default

- **Size of downloaded dataset files:** 1.97 MB
- **Size of the generated dataset:** 2.07 MB
- **Total amount of disk used:** 4.05 MB

**Fig 1.5 Size of Default Dataset**

In emotion case,

**Text:** a string feature

**Label**: a string feature

## Data Splits

| name | train | validation | test |
|---|---|---|---|
| default | 16000 | 2000 | 2000 |
| emotion | 16000 | 2000 | 2000 |

**Fig. 1.6 Data Splits**

### emotion

- **Size of downloaded dataset files:** 1.97 MB
- **Size of the generated dataset:** 2.09 MB
- **Total amount of disk used:** 4.06 MB

**Fig 1.7 Size of Emotion Dataset**

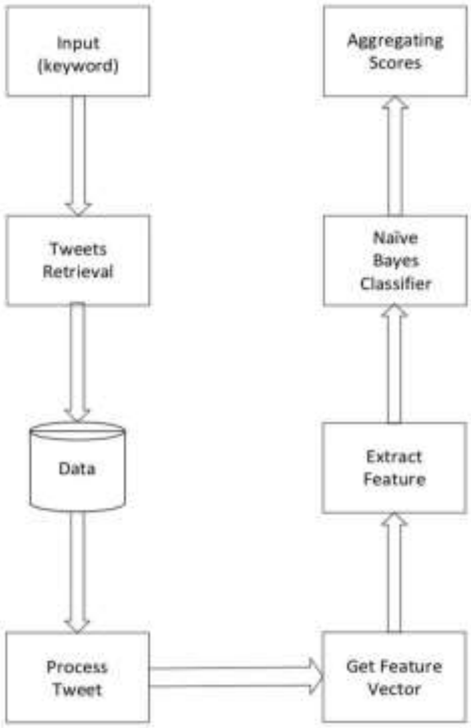


**Fig 1.8 Size of Flow Diagram 3**

**Wolfram Dataset**
[15]The Wolfram Data Repository offers a standardized method for storing data and instantly rendering it useful. The Wolfram Data Repository is a resource for global data that is easily available and it has been a hassle-free experience utilizing this dataset. It comprises datasets of all types and from many sources.

Our mental analysis model requires us to label each post as a positive or a negative emotion involved behind the social media post.

The Wolfram Data Repository makes use of the dataset after the preprocessing is done and immediately awards each post as a positive or negative emotion.

For Example,

```
RandomSample[ResourceData["Sample Data: Movie Review Sentence Polarity"],
 5]
```

| Review | Sentiment |
|---|---|
| no such thing is sort of a minimalist beauty and the beast , but in this case th ⋯ | negative |
| not for everyone , but for those with whom it will connect , it's a nice departu ⋯ | positive |
| it virtually defines a comedy that's strongly mediocre , with funny bits surfaci ⋯ | negative |
| meandering and confusing . | negative |
| if you can keep your eyes open amid all the blood and gore , you'll see del toro ⋯ | positive |

**Fig 1.9 WolfRam Dataset**

**The size of the dataset used in our model is 1346 Kb.**

VI. RESULTS AND FINDING
The past ten years have seen a rise in research into social media, and it is clear how important emotions are to how we communicate online from the quantity of research that is being done in this field. We have gained a great deal of information from the fieldwork that has already been done, and we can now ask ourselves how we might improve even further. We offer three specific

recommendations based on our review of the literature: employing more theories of emotion to assist the research, using more precise terminology, and taking into account whether distinct emotions provide better explanations than bipolar feelings. One of the interesting findings from the analysis of the literature was that although IS researchers are accustomed to borrowing concepts from other fields, it doesn't seem to be a frequent practice when discussing emotions in the context of social media. Although emotions have been actively explored within psychology for a long time, the use of theories explaining affect in the papers analyzed was minimal. Less than 10% of the articles employed a theory on emotion to influence their study questions or hypotheses. It would be intriguing to investigate why such theories are not applied more frequently. Is it possible that the majority of study to date on how people express their emotions online has concentrated on documenting what happens rather than seeking to understand why it occurs? Although they may not be deemed required for merely describing data, theories on emotion provide a solid foundation for thinking and interpreting observed behavior. Sentiment, emotion, and opinion distinctions It is difficult for social media researchers to use proper language because the terms affect, emotion, and mood are not easily distinguished from one another, and even psychologists have differing opinions on how to define them. We would like to suggest that one specific instance of ambiguous term usage that does not necessitate in-depth knowledge of the psychology of emotions deserves some attention. There seems to be an unstated presumption that sentiment and opinion are identical terms. Thoughts or feelings can both be referred to as sentiment. The same instruments might work well for measuring any of these at times, and both can be worthwhile research topics. However, we should be more specific about which one is being discussed when we present findings. Positive (or negative) opinions about something may not always correspond to positive (or bad) emotions that have been felt; in certain cases, they may even be the opposite. Consider a hotel review that says, "I'm pleased they're out of business!" as an example. Although the sentiment or emotion could be pleasant, the opinion is most definitely not. Opinion is something we're interested in if we want to know how highly individuals appreciate a service or good. Emotion is likely to be of greater importance if we wish to understand the motivations behind people's behavior and communication. It is probably not always accurate to generalize what we understand about opinions to emotions, or the other way around. We would want to propose that these two be clearly distinguished from one another and treated as two different entities when reporting findings. Differentiated Emotion to Bipolar, Another finding from the literature is that, up until now, most sentiment analysis has taken place on a bipolar scale. Recent research suggests that studying diverse emotions instead of just valence can provide us with more knowledge. We are aware that an emotion's level of activation affects the kind of behaviors it can elicit. For example, rage, a negative emotion with a high activation level, produces responses significantly different from melancholy, a negative emotion with a low activation level. It is possible that a finer-grained distinction between emotions than before would improve our comprehension of the events we study. Investigating, for instance, if an analysis utilizing distinct emotions may account for the discrepancies between the data in the Information dissemination category relating retweeting behavior and emotions would be intriguing. Why then don't we examine differentiated emotions more? A bipolar analysis approach may be suitable for the study's objectives in specific situations. It's also likely that despite some discoveries pointing in that direction, our field is still unaware of the importance of differentiated emotion. The fact that there are a lot more tools readily available for bipolar than differentiated sentiment analysis, or that these techniques are well recognized to researchers, is another potential contributing factor. It's important to remember that differentiating emotions may have varied means of expression depending on the situation or culture.

We have used 4 different combination of classification models like TF-IDF + Multinomial Naive-Bayes having the accuracy of 64%, CountVectorizer + Multinomial Naive-Bayes which has the accuracy of 74%, CountVectoriser + SVM which produced the accuracy of 84% and there best of the bunch TF-IDF+Multinomial Naive-Bayes which produced 85% accuracy.

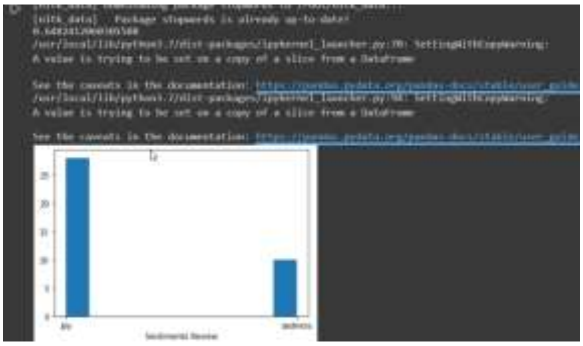

**Fig 1.4 TF-IDF + Multinomial Naive Bayes**

The above figure gives the accuracy of 64% which is just able to detect two different emotion such as joy and sadness

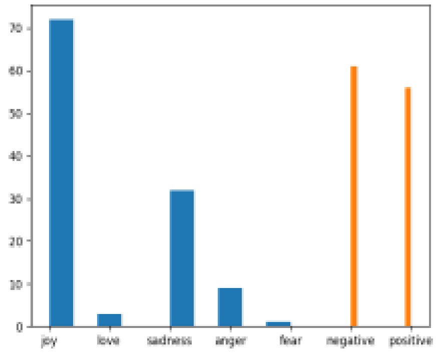


**Fig 1.5 TF-IDF + SVM**

The above graph is from the final Model which is of 85% accuracy with all the emotion and their polarity detected.

We tried our final model on many famous personalities with the likes of Elon Musk (CEO of Tesla and SpaceX)), with the final result giving Optimistics with the best mental till today i.e 89/100.

For finding out the results we have to compare our models according to their accuracy. TF-IDF+SVM proved out to be most promising and gave most accurate results with the accuracy of 85%.This model gave us the information about the classification of the tweets in order of different emotions which we took under the account of our research schema which involved considering a total of six emotions [anger,fear,joy,love,sadness,surprise]. All the emotions were given an arbitrary constant for scoring purposes.

The scores were as follows-

**Joy**-6, **Love**-5, **Surprise**-4, **Anger**-3, **Fear**-2, **Sadness**-1.

We also defined the twitter profile's polarity, either his/her account is towards the positive side or negative side. This gave us a different perspective from all the previous findings and researches which have been done before.

Each tweet gets a score and a user with multiple tweets gets a certain score for every tweet. We use statistical analysis of the scores we got from the tweets. And after integrating it with the polarity model we came to a conclusion about the overall result with score and polarity. Just like how the blood group works with a blood type and rhesus antigen which gives blood groups like A+. Similarly in our software model we see that and score come out with a score and polarity and according to some research and overall social media and psychological expertise we got from a psychology student which we referred from we got an overall idea about how these scores will map to certain psychological disorders which a user might have. These models give us accuracy about the tweets but not whether the mapping is accurate or not is still under review and may go through certain enhancements in future. This project gives a certain precautionary circumstance of probable mental disorder a user under view can have according to their tweets which we are examining. Currently we are using free twitter API for fetching a maximum upto 500 tweets thus giving a certain numerical and datawise disadvantage if the user is very active on his twitter account and has more than a few hundred tweets.

Examining the numerical scores with the mapping gives a certain mental health value which is in the form of a mental condition a user might have.

All the posts are taken into consideration and each individual gets a score and an emotion with a positive or negative value according to which the person is mapped to his/her mental condition.

A person could be anything. He could be **claustrophobic** or could have a healthy mind. He could be **jovial** or be in **delirium**. Everything depends on the score and the emotions attached with their posts.

People with bipolar experience both episodes of severe depression, and episodes of mania – overwhelming joy, excitement or happiness, huge energy, a reduced need for sleep, and reduced inhibitions.

Thus we come to a conclusion that if a person who has a mental score less than 1.5, referred from [20], which means that his tweets mainly consists of sad emotions .If the emotions are sad from his tweets we check whether the words he used is positive or negative if he uses positive words it can mean that he is instinctively sad but is hiding his emotions by using positive words which can be an result of his bipolar personality.

If there is a person with the same scores which is less than 1.5 and he uses negative words it can mean that he is not even trying to hide his depression, he can be depressed or just being sarcastic but as our model does not yet identify sarcasm we can only vouch for depression.

Person with a mental score between 1.5 and 2 means his tweets mainly consist of sad and fearful tweets which means that person is having an anxious personality he is afraid of clearly expressing himself when his other emotions are bound to take over him. He likes to be in his shell until something bad happens. So we placed those people in anxious personalities.

Person with a mental score between 1.5 and 2 and negative polarity means he does not particularly likes being positive towards the society he wants to criticize the social problems without acknowledging all the good things in it which explains that the particular person likes being in his space and despises the society because of his fear towards it.These people don't want to be excluded as well so they give their opinion.

Thus those people can also be afraid of being alone. These people can be categorized as claustrophobic or Socio phobic.

People with a mental score between 2 and 2.5 means his tweets are categorically angry and fearful either he is in a bad situation all the time or he is of compulsive personality who gets stuck on negativity all around him but his polarity also says a lot if he uses positive words he is being sarcastically angry and does not fear in expressing himself but he maintains a decorum while updating his social media so this says he is compulsive but thinks before writing anything but person with negative polarity and such score says he is being angry and fearful at the Behaviour of the society about almost everything he is being deprived of something he is being compulsive about his situation and does not want think twice before saying anything.

People with a mental score between 2.5 and 3 says that his tweets have a mixture mainly of sad, fearful, angry and surprised emotions he is full of negative emotions but if he uses positive words he is aware of his situation but is not able to adjust to it so he expresses himself in his way of constructively criticize things around him. But people with a negative polarity are not properly able to accept any kind of change and he is pretty vocal about it even in his wordings he is having an orthodox personality who is bad with change.

Person with a mid-score suggests the person is in control of all his emotions. He is equally happy, sad,angry,fearful etc he is having a healthy mind.

Person with more happy and surprised tweets with positive wordings has a jovial personality. He is happy and cheerful and if he is using negative words to define his happiness he is kind of delirious.

A person with a very high score means his tweets are generally of joy,love and surprised emotion. He is happy with what is happening around him and he is optimistic in nature. And if he is of negative polarity but overall his emotions are happy he is being calculative and does want to express himself fully he is calculative in nature.

VII. CONCLUSION

Since there are numerous machine learning approaches accessible, it is crucial to compare them all and then choose the one that best fits the target domain. Today, there are numerous specialized programmes in the medical profession that can forecast disease quite accurately in advance, allowing for effective and quick therapy. In the proposed work, five different machine learning approaches that were employed to categorize a dataset of different mental health issues were compared. The results make it quite evident that all five machine learning methods produce more accurate outcomes. All of the classifiers had accuracy rates of more than 79 percent. The research only employed a relatively small data set; in the future, a larger data set may be used, and the research could then be applied to it for more accuracy.

The studies and research that have already been done indicate that machine learning can be a helpful tool in understanding psychiatric diseases. In addition, it might aid in identifying and categorizing patients' mental health issues for future therapy. Modern methods that recognize patient responses and mood states among other things by utilizing data generated by the integration of numerous sensor modalities in high-tech gadgets have become increasingly popular.

It is apparent that most research and studies still have difficulty validating their findings due to a lack of sufficient, validated evidence, particularly from outside sources. In addition, not all machine learning techniques will perform equally well on all issues. The machine's performance Depending on the data samples obtained and the characteristics of the data, different learning models will be used. Additionally, preparatory tasks like data cleaning and parameter tuning might have an impact on machine learning models in order to get the best outcomes.

Therefore, it is crucial for researchers to explore and analyze the data using different machine learning algorithms in order to select the one with the highest accuracy. Additionally, difficulties and restrictions encountered by the researchers must be properly controlled in order to provide data that will be satisfying and will advance clinical practice and decision-making.

Depending on the data samples obtained and the characteristics of the data, different learning models will be used. Additionally, preparatory tasks like data cleaning and parameter tuning might have an impact on machine learning models in order to get the best outcomes.

Therefore, it is crucial for researchers to explore and analyze the data using different machine learning algorithms in order to select the one with the highest accuracy. Additionally, difficulties and restrictions encountered by the researchers must be properly controlled in order to provide data that will be satisfying and will advance clinical practice and decision-making.

VIII. FUTURE ENHANCEMENT

The proposed system can get more accurate with the results if the system used is of high performance. The system can have high accuracy and be less time consuming. We can provide more insights to the user about the emotions and the mapping. More models can be used to have more precise data. In Future, a lot more emotions can be used to map the person in a better manner. More analysis can be done using psychology research papers to learn more about mental health issues. We could take help from neurosurgeons to have more knowledge about the brain so that we can provide more details about the health issues to the user so that the analysis doesn't seem superficial. Working with trials to research more about the problems so that we can tell a user how to go about their problems and how they can prevent other issues from happening. The practice of psychiatry is anticipated to become more individualized in the future and science and psychiatry need to work together to treat mental health issues in an efficient manner.

IX. REFERENCES

1. THE IMPACT OF ARTIFICIAL INTELLIGENCE ON INNOVATION Iain M. Cockburn Rebecca Henderson Scott Stern NATIONAL BUREAU OF ECONOMIC RESEARCH 1050 Massachusetts Avenue Cambridge, MA 02138 March 2018
2. A survey on feature selection methods, by Chandrashekar, G., & Sahin, F. Int. J. on Computers & Electrical Engineering
3. A survey on natural language processing (nlp) and applications in insurance Antoine Ly, Benno Uthayasooriyar, Tingting Wang Research on Principle and Application of Maximum Entropy, Boya Wu; Junyan Yi; Qiaoling Yong
4. Bayesian Multinomial Naïve Bayes Classifier to Text Classification, Conference: International Conference on Multimedia and Ubiquitous Engineering International Conference on Future Information Technology. Shuo Xu, Yan Li.Wang Zheng
5. Research of the Text Data Vectorization and Classification Algorithms of Machine Learning, June 2020, Theoretical & Applied Science, Vadim Kozhevnikov, Evgeniya Pankratova
6. Research paper classification systems based on TF-IDF and LDA schemes, Sang-Woon Kim & Joon-Min Gil
7. Using NLTK for educational and scientific purposes, Mykhailo Lobur; Andriy Romanyuk; Mariana Romanyshyn
8. A comprehensive survey on support vector machines in data mining tasks: Applications & challenges,Janmenjoy Nayak, Bighnaraj Naik,Dr. H. S. Behera
9. Scikit-learn: Machine Learning in Python, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay
10. The NumPy Array: A Structure for Efficient Numerical Computation, Stéfan Johann van der Walt, S. Chris Colbert, Gael Varoquaux
11. pandas: a Foundational Python Library for Data Analysis and Statistics Wes McKinney]
12. https://www.sciencedirect.com/science/article/pii/S1877050919311561#!
13. https://huggingface.co/datasets/go_emotions
14. https://datarepository.wolframcloud.com/
15. https://www.sciencedirect.com/science/article/pii/S2090447914000550#!
16. https://dl.acm.org/doi/abs/10.1145/2436256.2436274#pill-authors__contentcon
17. Sentiment Analysis of Twitter Data Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau
18. Exploration of Artificial Intelligence for emotion recognition using physiological signals by Arpit Bhardwaj, Prabhjot Singh and Pradeep Tomar
19. The abbreviated mental test score(AMTS) by Hodkinson in 1972.