Prediction of Mental Health among Twitter users

Research Project

MSc in Data Analytics

Naval Suvarna

X18183654

School of Computing

National College of Ireland

Submitted to: Dr. Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Naval Suvarna		
Student ID:	X18183654		
Programme:	MSc In Data Analytics	**Year:**	2019-2020
Module:	Research Project		
Supervisor:	Dr. Vladimir Milosavljevic		
Submission Due Date:	17/08/2020		
Project Title:	Prediction of Mental Health among Twitter users.		
Word Count:	12808	**Page Count:**	30

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …………………Naval Suvarna……………………………

Date: ……………..16/08/2020…………………………………………………….

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Mental Health among Twitter users

Naval Suvarna

X18183654

Abstract

Awareness about Mental health has been on the rise in the past few years as opposed to the earlier ages when it was neglected due to lack of physical evidence and the fear of being stigmatized. Along with one's physical health, mental health is also crucial to an individual's well-being. Sometimes this mental health can be affected due to life-style changes or circumstances that induce a life-altering feeling. When mental health is compromised it could lead to the deterioration of an individual's overall-health. Signs of mental distress can often be found on the online activities of individuals via exploration of posts that have been uploaded online. Since traditional diagnosis take place after the condition has turned worse, there is a need for a quick yet accurate system which would provide early warnings about the condition of such individuals. This research has been carried out to predict mental health of Twitter users using machine learning models and deep learning and also to recognize the optimum model that could be used on such data. The research culminates by recognizing that MLP classifier and Random forest classifier are the most optimum models that can be used to predict the mental health of individuals using tweets.

1.Introduction:

The human body has many vital organs functioning in harmony continuously carrying out voluntary and involuntary tasks. However no other organ other than the brain is more crucial in carrying out the proper functioning and daily metabolic activities that ensures our survival. Although the brain is such an essential organ on which our life depends, we seldom take care of its health and well-being. Our psychological and emotional well-being is product of how well our brain functions. This is also known as mental health (Bhaugra, et al., 2013). Mental health is an integral and crucial part of overall health, and can be explained being in a balanced-state within oneself. Also, the state of being mentally healthy means that the person can build and sustain emotional bonds with other beings, and can participate in social activities and responsibilities in their culture along with identifying and acknowledging emotions and feelings such as happiness or sadness. The absence of this integral functioning is termed as mental illness. Mental Illness like mental health can be affected by a number of social, biological and psychological factors. Experts suggest that mental health is vulnerable due to internal factors ranging from lacking emotional resilience to poor social status and isolation (Bhaugra, et al., 2013). In simpler terms, mental health can be described as an individual's thought process concerning themselves and their lives. Due to the role mental health plays in the crucial functioning of one's overall health, it should not be neglected.

1.1Background and Motivation

Instances of self-harm has been on the rise in the past few decades. Testimonials from these people indicate that they suffer from mental illness. However, since mental illness cannot be physically quantified or the symptoms are not common such as headache or common cold, a majority of the population is still ignorant towards this condition. This ignorance towards the condition has led to many incidents where mentally ill individuals fight the condition on their own or are completely dominated by it which leads when towards extremes steps such as suicide or homicide (Vatrapu, et al., 2016). Also, the stigma that is attached to mental illness makes it all the more difficult for people who suffer from it to express themselves (Vatrapu, et al., 2016). It is observed that human are a social tribe and we all have the need to connect and to be valued. Since the condition that individuals go through while mentally ill, this decreases their self-esteem and self-worth. It makes it difficult for them to interact with other humans. This causes isolation and the individual shut themselves up totally avoiding any social contact which worsens their conditions but the need to feel valued and cared for is still present. This causes many to turn towards social media platforms which gives them an alternative to social contact without the risks of being exposed or judged. Social media also provides a platform for expressing their thoughts, their views and their emotions under the cover of anonymity. Also, research has proved that online behavior is influenced by mental health. Mining this data can help us uncover clues in online activity that affects mental health. Advancement in tools that can process textual information and draw insights from them can help in this process. These tools focus on the text and after its processing decide its polarity. It is often said that language is a straight-forward implementation of a person's thought process. Therefore, processing textual data can lead to a number of new possibilities in understanding of mental health. Natural language processing is been used in almost every field of textual analysis. NLP is the branch of artificial intelligence which is responsible for dealing with human interaction with computer through the use of natural language. In simple terms NLP reads and deciphers the human languages to understand and make sense of it. Easy availability of data coupled with advancements in the computational abilities has expanded the possibilities of these toolkits to better understand and interpret human languages, leading to a surge in their applications in various fields ranging from retail to security. Classification is the most common used technique in NLP which extracts cues and frequents words based on a matrix of their occurrence in data (Shen, et al., 2017). Since Facebook and Twitter have a dominant share in social platforms where users upload their views and express their emotion, NLP could be applied on these platforms (Shen, et al., 2017). However, the major shortcoming of NLP is that though it can process, decipher and even interpret human language it cannot carry out predictions based on those interpretations. Another field of artificial intelligences is machine learning which has exhibited great promise for implementing predictions based on interpretations. Using machine learning models such as deep learning can help analyze the textual data from tweets and post on social media. Since research suggest that people are expressive on social media and that mentally disturbed individuals express their emotions more openly on the internet, the idea of implementing machine leaning models on social media data could help identify individuals who are mentally disturbed. In this research, Twitter has been selected as the platform since it is the most text-rich source among other platforms and has a limit of 140-characters. Other platforms have no such limits on the number of characters and often are accompanied by videos or images that are of no use for text-analysis tools. Since twitter contains a rich amount of textual data it has been chosen as the preferred platform for predicting mental illness.

The main goal for this research project is implementing the most efficient machine learning model which can predict the mental health of twitter users by analyzing their tweets. The implementation of this project will help identify individuals who have mental health illness. In order to carry out this implementation, a number of machine learning models will be applied to tweets that has been derived from the Twitter dataset from Kaggle[1] repository. The main attributes on which models will be applied are the tweets. In this dataset sentiments of the tweets are previously established. Mental health will be determined based on the timestamp of the tweets and the tweets themselves. The models will predict the nature of mental health based on the tweets. The attribute timestamp has been taken into consideration since most researches mentioned in Section 2 have suggested that mentally ill individuals suffer from insomnia and disturbed sleep and are more likely to express themselves during those periods. The model implementation has been explained briefly in Section 5 and the corresponding results have been detailed in Section 7.

2. Related Work

2.1 Learning about Mental Health

According to (Prince, et al., 2007), neuropsychiatric disorders attribute to about 14% for the burden of diseases globally. The reason for this is mostly the nature of depression along with other mental disorders bordering on alcohol-use and substance-use which are chronically disabling. These estimates are a cause of concern and have drawn the attention towards mental health. There has been alienation of mental health from the primary efforts towards improvement of health due to separate contributions of mental and psychical disorders. Mental health has been neglected due to inadequate understanding of the connection mental health has on other health conditions. (Prince, et al., 2007)claims that the interaction are so versatile there can be on improvement on overall health without first understanding the complex nature of mental health. Mental disorders have proven to proportionally affect communicable as well as non-communicable diseases which lead to degradation of the immune system. The effects of mental health and overall health are interdependent on each other, as many health conditions increase the likeliness of mental disorders among patients. Since mental illness cannot be physically quantified, health services are often not provided equitably to people with underlying mental illness. (Prince, et al., 2007)calls for a plan to “develop and evaluate psychosocial interventions” which will them be integrated into management of over-all diseases. Health care apparatus need to be revamped on a global level to accommodate the need for mental illness which is on a rise due to the current lifestyles and socio-economic events. Since stress in on the rise, so is the occurrence of mental illness. This gives an opportunity for tools which can detect and identify the likeliness of mental illness to start developing.

For better awareness of mental health, we need to understand it properly. (Westerhof & Keyes, 2010) defines mental health as the state of having no psychopathologies which include depression and anxiety. Lifespan development can benefit from the absence of mental illness. (Westerhof & Keyes, 2010) focuses on three core components of positive mental health such as social well-being, psychological well-being and emotional well-being. Emotional well-being is the presence of feeling happy and satisfied with life. Social well-being is defined by the positive feeling of being socially functional or being high in ranks of social value.

[1] https://www.kaggle.com/

Psychological well-being is referred to the positive individual functional or a sense of being happy in terms of self-realization. (Westerhof & Keyes, 2010) has also tried to understand the relation of age and mental health and illness. Models such as regression analysis point to the fact that mental illness is prominent among the older adults, but are next to none among today's oldest-old. Mental health starts to deteriorate among the age group of 40's. (Westerhof & Keyes, 2010)concludes that episodes of mental illness are experienced very less by the oldest age group. The issues of mental illness among the older adults can be attributed to life circumstances which are age-related. Changes in lifestyles, added responsibilities, work-related stress or the usage of social media are also some of the factors that were discovered during the studies.

(Corrigan, et al., 2014)sheds light on the stigma that comes from being mentally unstable. The social stigma arises from an inadequate knowledge about mental health and how it affects a human being. The social stigma prevents people suffering from mental illness to ask for help or undergo treatments. They may even be unwilling to participate in group sessions that support their conditions. The perceived stigma that arises from being recognized as mentally ill also prevents individuals to identify for themselves the very onset of depression or anxiety. The fear of being socially isolated and being branded as mentally incapable further spirals the condition in the downward direction. (Corrigan, et al., 2014) identifies stigma as a complex construct which has public, self and structural parts. Stigma has shown to not only affect the ones with mental illness but also their provider networks, communal support and support systems. Knowledge and awareness about mental health will play an important part to moderate the effects of stigma. They identify two types of barriers which affect the sentiment towards mental health. Person-level barriers lead to neglecting the concern for mental health and in turn may affect health decisions. Lack of proper medical care, insurance, financial status consists of system-level barriers. (Corrigan, et al., 2014)also mentions the various symptoms of a serious mental illness. Lack of motivation and lethargic behavior, fading of interpersonal skills and the general lack of analytical and assertive thinking.

2.2 Social Media and Mental Health

The Connection between social media and mental health has been the topic of research worldwide in the last few decades. Researchers have been trying to find the effect social media has on mental health. (Lin, et al., 2016) has documented the research on the association of social media and depression among U.S Young Adults. A study was conducted among 1787 users and parameters such as total time spent per day and visits per week were studied. Depression was assessed using a depression scale based on Outcomes Measurement System which basically consisted a questionnaire. This questionnaire consisted of questions about the user's feelings and their reactions to certain events. They also recoded the mood of the users surveyed. Statistical models such as Chi-square and ordered logistic regression were used along with sample weights. The study reported strong relation between social media use and depression. It is also reported that individuals suffered from mood dysregulation after social media use. Some also reported using social media to express concerns about anxiety and depression. It was also discovered with the tests that depressed individuals who usually had a low sense of self-worth, often turned to social media for online validation. (Lin, et al., 2016) also states that due to the high availability of social media and the chance to socialize in a controlled setting attracts individuals with depression or anxiety to social media. This study concludes that though there is a linear relation between social media use and depression, it is also found that self-disclosures of depression are common on social media. (Lin, et al., 2016)mentions proactive efforts taken by various social media sites to reach out to users who have indicated sign of mental illness.

(De Choudhury, et al., 2013) states that depression can be a challenge to identify and so they have examined the potential of social media postings to understand depression in populations. Information that has been obtained from social media has potential to aid the research that is carried on with surveys to identify depression among people. The paper has used a methodology of using Twitter posts which are shared by individuals who have been previously diagnosed with depression. They also plan to initiate training of a model that can determine if certain posts are indicative of depression. The model will learn from different parameters such as social activity, language manifested and emotion. Using this model (De Choudhury, et al., 2013) aims to achieve the aim of introducing a depression index based on social media that can characterize levels of depression. SVM classifier has been used to help with the indication of depression. The models have been able to predict the indication of depression with an accuracy of more than 70% while a precision of 0.82 was received. SMDI (Social Media Depression Index) a metric is developed which could point out the depressive-indicative posting that were posted on Twitter. Features such as emotions, time of the posts, their linguistic style, and their language were all taken account. Users with depression tend to post using negative emotions at late hours of the night and use words that have negative sentiment in them. All these factors were used to train the models and achieve SMDI. The paper concludes by stating that the user’s behavioral data may play a key role in understanding the depression inducting incidences and may help with further research.

2.3 Using Social media as a tool

(De Choudhury, 2014) has documented their efforts towards using social media as a tool for predicting depression among Facebook users. Facebook is a social media which enables users to share various posts among a group of people with has been consented by the user. It also enables a user to like comment and share on their friend’s post. The research intends to characterize postpartum depression among these users. Various factors such as social capital, linguistic styles, activity are taken into consideration when using machine learning to detect. Also, data from pre-natal and post-natal periods are used to detect. It was found that social isolation coupled with lowered availability are the best predictors of postpartum disorders. The research also reports that a lot of cases of depression go unreported due to fear of stigma, and frequent mood swings among mothers. The model that used only pre-natal data is said to give 31% variance.

(Shen, et al., 2017)has used techniques to harvest social media as a means to detect depression. It is said to be the need of the hour since more than 70% of the patients had no idea about the onset of depression. It states that people have increasingly displayed emotions on social media which makes it a goldmine of emotional information. Well-labelled datasets have been extracted using twitter and six features have been leveraged to predict depression. A multimodal depression dictionary has been developed with the help of this model. The datasets have been put through an extensive series of data processing using the NLTK toolkit and WordNet. Emoji conversion, stemming and removal of stop words are most of the steps used.

Social Network Feature indicated that twitter was a way of emotional interaction for depressed users. Several such features were extracted from the datasets. Statistical models are used to detect the depression likeliness based on these features. It was found that depressed users use 0.37 positive words and 0.52 negative words on account of every tweet. This the paper concludes that by harvesting such data, depression and other mental illness can be predicted among users.

(Guntuku, et al., 2017)has attempted to detect mental illness on social media with an integrated review of different machine learning and statistical models which will aid the identification of symptoms. This study employs self-surveys, posts on Twitters, their style of language and online activity. Automated detection methods are employed to identify symptoms with mentally unstable subjects with the help of large-scale passive monitoring. Automated analysis can be achieved by training predictive models, which utilize features or values that social media data contain. The common features are the frequency of words, the times at which the posts were shared among other features. Linear Regression and SVM have been used as models to achieve the desired outcome. Cross validation has been used to prevent overfitting. The Linguistic and word count approach has been used to pick ques for characterizing ambiguity among several conditions. (Guntuku, et al., 2017) also claims that since mental illness carries social stigma and discrimination, data protection and privacy frameworks need to be strict in order to prevent any harm to those diagnosed. The paper conclude that various online variables need to be considered in order to detect depression or other mental illness among users. (Saravia, et al., 2016) proposes a novel technique that collects online data and leverages the power of predictive models which employ language and behavioral patterns, used specifically on tweets to determine the persons mental stability. The system known as MIDAS, is a web-service which characterizes a person's behavior and linguistic style. The predictive model is leveraged for implementing an analytics and assessment tool. This tool will have the capacity to provide statistics in the form of visual data and will have a functionality to assess the individual's mental health with the help of twitter data. (Saravia, et al., 2016) claims that this system is beneficial for those who are physically and mentally unstable to take tests or surveys for their conditions. They intend to use term frequency inverse document frequency which is a method to assign weights to words based on their occurrence in a sentence. The Pattern of Life feature will also be leveraged in order to better the results. The Twitter Ids have been sampled on the basis of features based on linguistic and behavior. Sentiment 140 extension has been used to label the tweets based on their sentiment and the polarity is then transformed into 5 different polarity. Overall (Saravia, et al., 2016) mentions the efforts required to build an online system for exploration of various properties. Also feedback from the patients is taken and implied to further improve the accuracy of the model. (Aldarwish & Ahmad, 2017)has implemented the use of Social Network Sites to study user-generated content in order to understand the onset of mental illness among online users. User generated content contains all the data about a user including their sleep patterns, thinking, mood, natural behavior, loneliness, worthlessness etc. The proposed tool is to be intended for use of medical professionals, family to track the onset of mental illness before it reaches a major phase. SVM and Naïve Bayes classifier is used to test seven major operation in two datasets. The models are trained on the first dataset which contains depressed and non-depressed posts. The second dataset contains SNS post and is used for prediction. The first operator is Select Attributes which does as the name suggests. The second and third operators are Nominal to text which changes the type of the selected attributes. Process Documents for the fourth and fifth attribute which is used to train the dataset and to generate word vectors. Validation forms the sixth operator which validates the model. Apply model is the seventh operator and connects the training and the testing dataset. Post are collected from twitter, Facebook and LiveJournal. The models achieve a satisfactory level of accuracy while prediction which gives way to the fact that multiple social media can be scrutinized for the detection of mental illness.

Apart from the field of mental illness, (De Choudhury, et al., 2013) social media proves a promising tool in order to predict the mood and behavior of mothers who have just gone through child-birth. Twitter posts have been used to collect data and quantify postpartum changes among new mothers. The data covers all dimensions such as linguistic style, social network, emotion and social engagement. The predictive models will be able to accurately point out the mother who will have a change in their behaviors. The study is motivated by a need for a privacy-sensitive behavior detection system which can study the changes in mood. The birth of a child is major change which leads to additional responsibilities often accompanied by changes in sleep and other routines which can cause stress and anxiety to pile up. Often hormonal changes are the scapegoats in these conditions, but the study is motivated to find out other reasons as well. Principal component analysis has been used to prevent any chances of over-fitting or feature-redundancy. Classification models such as Random Forest, Naïve Bayes, decision trees and SVM are used. SVM seems to over-perform with the highest accuracy. Five-Fold Validation has been used on the dataset in order to understand the possibility of model overfitting. The study concludes with a point that online activity can be harnessed for predictions on people's mood.

(Naslund, et al., 2016) proposes a model that creates an online peer-to-peer connectivity. This online system would be helpful for someone who is suffering from mental illness and needs a support. This support system could help individuals fight mental health and the stigma associated with it after they realize they are not alone. It can also help individual with online interventions about mind and the illness that ails them. This system will help them learn from their peers, talk about the issues that bother them and also provide a support to someone if need be. Privacy is one of the major concerns that comes across while implementing a platform such as this, since the profiles will be available to everyone in the forum and the possibility of leaks are high. The study reveals that young people with mental illness have a tendency to speak out on the internet through blogging and were likely to build friendships online where people could connect and share matters online. If the predictive models are successful in identifying profiles with genuine mental illness, this peer-to-peer system can have a use in the digital age for the fight against mental illness.

2.4 Using Twitter to detect mental illness

(A. & Sonawane, 2016)has provided a survey of techniques to comparatively analyze the techniques used for opinion mining. Different machine learning techniques such as lexicon-based approaches have been used in some studies along with evaluation metrics. Naïve bayes, Support Vector Machine and Max Entropy were the techniques covered for research in to data analyses of twitter data. It has also discussed the application of different techniques for sentiment analysis which often ends up being either positive or negative for twitter data. They usually use Textual information retrieval methods for processing, and analyzing the facts. The contents that often are processed include attitudes and emotion. Sentiment analysis of a data is nothing but automating mining of opinions and views from textual data. The data can be mathematically denoted as a matrix of vectors. They have developed a sentiment classifier which is primarily based on Naïve Bayes method. This method usually uses features such as N-gram and POS-Tags for predicting the sentiment of a given tweet. The tweets were analyzed by two models one of which is a bigram model and one with Maximum Entropy model. Different data cleaning techniques have been mentioned in the research. (A. & Sonawane, 2016) has used the Python NLTK toolkit for implementing the Naïve Bayes model. They have concluded with the fact that the success of the model depends on the selection and extraction of features for sentiment analysis.

(Husseini Orabi, et al., 2018)proposes a technique with word embedding optimization for classification of sentiments and identification of depression from tweets. The approach has been used to improve the performance of models already applied on two datasets for depression detection. The datasets are CLPsych2015 datasets and Bell Lets Talk Dataset. A comparative evaluation approach has been taken to investigate and the performance of deep learning models that have previously been used for depression detection. Word embedding and parameter hyper tuning has been carried out to ensure accurate results. As the twitter data is characterized as Control, PTSD and Depressed the training set has been divided accordingly. (Husseini Orabi, et al., 2018) has performed CNN and RNN- based models for prediction. Word embedding are generally used on unsupervised training data. They are then used to solve supervised tasks. Word2Vec is another model which has been used reconstruct a word context. CNN based max pooling models have reported better performance than RNN which concludes the analysis of this research.

(Houghton & Joinson, 2012)has concentrated its efforts on forecasting the onset of mental illness with data obtained by twitter with linguistic markers of confidential data uploaded on twitter. The novelty of this research is that the paper has also taken into account posts from Secret twitter which has all the properties of twitter but with an element of anonymity. The platforms were chosen since they have a limit for message length. This indirectly helps control breadth of disclosure and only measure a difference in depth. The LIWC software is used since it is a widely accepted linguistic tool and the advanced word counter feature distinguishes words in the tweets to different linguistic categories. It contains 80 categories which belong to the linguistic and psychological criteria but can only count words which can or cannot be contained in word stems to verify the distinguish features. Two Naïve raters were utilized in an attempt to ensure that the secret posts were more sensitive than the normal tweets. There were certain other distinguished markers among the tweets that were identified among the data. Normal tweets contained articles, pronouns in third person and swear words along with Exclamation marks and punctuations. Whereas secret tweets contained a lot of personal pronouns, human inhibitions and sexual context. The 16 predictors are usually enough to predict the onset of mental illness with satisfactory accuracy and strength.

(Kim, et al., 2017) has mentioned the various tedious tasks that complement the analysis of Twitter data. The first major impact is the sheer volume of tweets that need to be studied. They contain millions of cases and many traditional software packages do not have the capacity to handle them. It then becomes a mundane task of exporting the data in batches and then assembling them into working datasets. They have chosen a dataset of more than a million tweets. They choose Radian 6 export tool which maxed out at 5000 cases, which resulted in multiple file exports and then imported into Text analytical software. There are other softwares which can handle large amount of data but they need high amounts of investment and the learning curve may not be suitable for single studies. Data cleaning which is an absolute important step for text analysis needs to be handled with upmost care. This can be a crucial step which ensures the accuracy of the data and the performance of the model. Text cleaning is said to be an expensive step. (Kim, et al., 2017) has mentioned that the analysis of text data is hampered due to use of slang and unconventional slang of writing such as sarcasm, hashtags and emoticons which use jargonized text to convey feelings. For the above reasons, automated software packages are highly insufficient when used on tweets. Manual codding is highly advisable for twitter data but it is high cost-effective and hence is avoided. Statistical tools and predictive models are used in unison to extract patterns and meanings. Also, many commercial vendors sell automated codes but are seldom customizable and are a burden on the budget.

(Coppersmith, et al., 2015)has tried an attempt to quantify the mental health signals on twitter, with the help of rigorous application of language processing methods. They have proposed a novel method to collect data for a wide range of mental illness, particularly four which includes PTSD, bipolar disorder, seasonal affective disorder and depression. The study of mental health research pre-requisites the study of behavior, how a person behaves, the way that they communicate, the types of activities they choose to engage in which forces their interaction with people. The LIWC has been used here as well as a tool to quantify psychological behaviors. Features have been extracted for subsequent machine learning models and the user's language and behavior has been quantified. Via automated models. Language models have been used here to predict the likeliness of a word sequence. N-grams have been eliminated as a viable option for mining social media data because of inconsistencies and noise in the data. A 1-gram LM has been implemented which would predict the likeliness of a whole word. After that a 5-gram language model has been applied for sequences of 5 characters. The fact that insomnia and sleep disturbance has been the symptoms of mental illness, the tweets that have been recorded from midnight to 4 in the morning at local time have been scrutinized. Also, negative and positive sentiments have been taken into account as features for machine learning models to study the mental health of users.

2.5 Ethical Complications of using Social Media

(Joseph, et al., 2015) has portrayed a detailed outline of the misuse of certain words that relate to schizophrenia which are generally misunderstood by people on the internet. The paper explores the probability of stigma not only in the outside world but also people being bullied due to the use of certain words related to schizophrenia which they have posted online. Use and misuse of tweets published on the internet is a matter of concern for many individuals. Data privacy is a matter of strict regulations and have attracted serious penalties for not being satisfactory. Government authorities in certain parts of the world especially US and Europe have placed strict guidelines and regulations for sharing personal data that have been published on the internet. (Joseph, et al., 2015) has also examined the misuse of the words that have a relation to mental illness with other health conditions. Chi-square tests were used in this venture to identify the proportion of misuse among twitter users with regard to other mental health. The use has also been compared through a series of different criteria like sarcasm, negativity, not a reference and inappropriate. These levels help us decide the proper or improper use of the term schizophrenia.

In lines with personal data markets, (Spiekermann, et al., 2015) has outlined several challenges which have been faced in the research of data analysis that revolve around the data privacy domain. It has defined personal data as a new asset class, which is also an ecosystem for collecting, analysing and transacting personal data. It is also seen an added value for businesses and companies. Personal data has a lot of customers in the open market. Thus, the management of personal data becomes a liability. Privacy regulation is an evolving field of law with a main focus of protection of personal data. The liabilities that arise in the management of data is mainly cybercrime and identification of financial information. Organizations which fall prey to these violations have to go through strenuous procedures to regain the licence to hosting data again. The GDPR in Europe deals with the management and storage of data and also its transfer outside of Europe for purposes of archival or processing. has detailed a lot of challenges that companies hosting data face due to these necessary regulations.

The work of publishing the findings about twitter is of sensitive nature. Although twitter is an open platform, a sufficient ethical justification to collect and process that data needs to be provided. (Webb, et al., 2017) reports on the best practices that should be followed and an ethics consultation that should be meet to prior to every research that is being undertaken. (Webb, et al., 2017) have documented the challenges they faced while handling data and processing them. Most of them related to the private nature of the tweets being published in a research dissemination and the question of whether it was necessary to contact the original users whose tweets were being analysed. It is absolutely crucial to solicit their informed consent for leaving very low scope of any action. The authors have consulted ethics guidelines and have adhered to expert opinions on these matters. They have reported that the publication of Twitter data raises concerns of significance in terms of ethics but there is no absolute solution on the matter as to how it needs to be resolved. There may be times when good practices have compromised academic integrity. Concerns about privacy in the project were covered by following guidelines such as anonymising the usernames and @ handles. The Digital Wildfire project screens tweets for hate speech. Other issues such as an improper presence of academic consensus, and informed consent have been discussed in length here. The research concludes that though twitter is an open platform the right to use it for private or public research is debatable and a precise well guided regulation is needed.

3.Research Methodology

For the achieving the goal of this project various methodologies have been compared to best suit the implementation. Cross Industry Standard Process for Data Mining also known as CRISP-DM is used here since it is an open standard and has been widely used for similar projects in text analysis.

3.1 Data Set Overview:

A single dataset has been used for this project. It has been sourced from the Kaggle online repository and is publicly available.

Twitter dataset: It contains 1.6 million tweets for a period of one year. The features available in the dataset are as follows:

Target: This attribute contains the sentiment of the tweets. 0 is labelled for negative sentiment, 2 for a neutral sentiment and 4 is designated for a positive sentiment. The targets are predefined in the dataset.

Ids: This field contains a distinct number for the tweet.

Date: This field contains the timestamp of the tweet. Time of the day along with the day ,month and year has been recorded.

User: This field mentions the twitter handle of the user.

Text: This attribute contains the text that is to be analyzed. It contains the tweets posted by the users. The text will be used as a feature to analyze mental health.

The dataset contains an equal number of tweets with positive and negative sentiment as shown by the below graph. There are no tweets with neural sentiments.

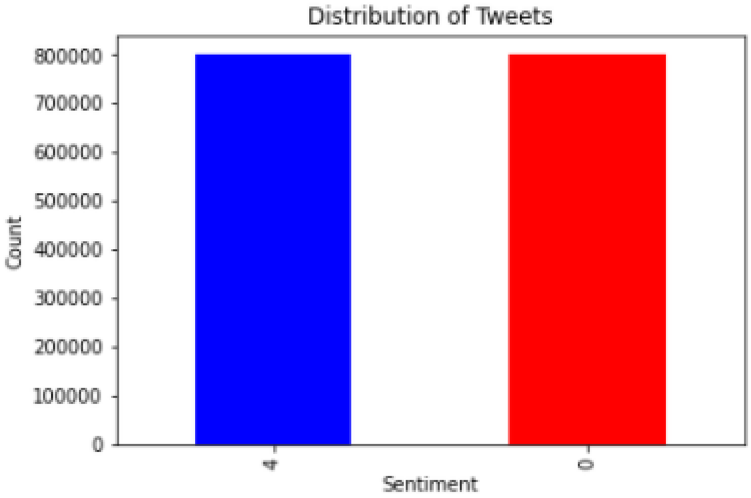

Fig 1. Distribution od positive and negative tweets

3.2 Data Preparation

This is one of the vital steps for working with data and models. The accuracy and precision of a model can be affected due to substandard data being passed into the model. The model requires only a certain fraction of features to perform the predictions that have direct correlation with the target variable.

Since this research deals with text a lot of steps need to be followed in order to make the natural human spoken language such English into a machine interpretable language. There are only a few attributes in the dataset that are of use to the model. Those features that have no relevance to the model need to be discarded. Preparing the data to be interpretable for a computer involves steps such as data cleaning and data pre-processing. Since online posted tweets contain a lot of special characters and number or emoticons, they need to be removed. Modelling the data with such irrelevant parts tends to decrease the accuracy of the data. Common words or preposition need to be filtered out. Data pre-processing contains steps that transforms the data to be interpretable by the model. This includes vectorizing the data which assigns tokens to the words based on their frequency. Also, any feature addition needs to be executed in this step. If the data is skewed towards one target variable, resampling techniques need to be carried out so that the data sample is balanced since it can affect the precision of the model. The data needs to be split into train and test set using different splitting techniques.

3.3 Modelling:

Modelling the data is a crucial step in CRISP methodology which enables the model to learn from the clean data that has been split into training and testing sections. A total of 5 Classification models have been applied on the dataset.

3.3.1 SVM:

Support Vector Machines are known for their ability to solve linear and non-linear classification as well as regression problems. The algorithm classifies data by the idea of a line or a hyperplane that is created which then separates the data into distinct classes. The algorithm separates the data based on multiple possibility of classifying them and then chooses the best fit classification hyperplane among them. When the classes are separated by the hyperplane, the closest point of data to the hyperplane on either side is known as support vector. The hyperplane which has the maximum distance between the two support vectors is the optimal hyperplane. This selection among multiple possibilities helps SVM build a robust model able to classify accurately.

The model has been used in text analysis research previously and has a strong standing in terms of performance. (A. & Sonawane, 2016) have used Support Vector Classifier along with other models to predict mental illness using social media posts and has out-performed other models. Similar attempts to use SVM have been carried out in (De Choudhury, et al., 2013) and (Aldarwish & Ahmad, 2017) where the primary field of analysis were social media post.

3.3.2 Naive Bayes:

A probabilistic machine learning model, Naïve Bayes is generally used to solve classification tasks. The algorithm is primarily based on Bayes Theorem which works on the probability of events occurring given a set of conditions. There are different types of Naïve Bayes algorithms used such as binomial, Bernoulli's and Gaussian Naïve Bayes. Gaussian Naïve Bayes has been used in research involving sentiment analysis, recommendation systems and spam filtering. It is widely used due to its ease of implementation and shorter training time as compared to several other models.

Naïve Bayes has been used in (A. & Sonawane, 2016)to classify and predict social media posts labelled as depressed and non-depressed using seven major features. Also used in (Aldarwish & Ahmad, 2017) and (De Choudhury, et al., 2013) has played the role of a sentiment classifier in similar researches.

3.3.3 Random Forest:

A supervised learning algorithm, it builds 'forests' which are an ensemble of decision tress. They are generally trained with the bagging method. The bagging method is effective since the combination of learning models has a positive effect on the overall result. In simple terms, random forest creates multiple decision tress and combines them together to obtain a more stable and accurate prediction. The performance of random forest also depends on the fact that it introduces additional randomness while growing the trees.

An approach towards using random forest has been taken in (De Choudhury, et al., 2013) , where other machine learning models have also been used to predict depression among social media users. It has been mentioned that random forest has a better performance than decision tress, and lower learning time than SVM.

3.3.4 XGBoost:

Since tree boosting algorithms perform well, XGBoost which is an ensemble technique takes the performance up a notch due to its application of gradient boosted decision trees.

XGBoost uses a combination of weak learners that have predicted the output incorrectly and uses it as a feedback to rectify the errors. The weak learners are then combined into strong learners which is the reason for its higher accuracy rates. The package is said to be highly scalable to larger datasets and can be optimized with a large set of parameters to obtain efficient computational performance while handling sparse data. Since the dataset contains a sparse matrix after being vectorized, this algorithm has been chosen to effectively analyze and predict the outcome.

3.3.5 MLP Classifier:

Multi-layer Perceptron Classifier is a module in the Scikit-Learn toolkit which performs the task of classification for which it relies on neural network. This makes it different than the rest of the classification algorithms used. A perceptron works as a linear classifier, which classifies input by isolating two categories with a line. Multiple layers of perceptron have the ability to classify non-linear functions as well by approximating an XOR operator. Multilayer perceptron are the basic implementation of deep artificial neural networks. MLP classifier are generally used for supervised learning problems and are feedforward networks. This means that every prediction is performed from the learning and used as a feedback to improve on the performance of the networks. The forward pass mechanism implies that the signal flows from the input layer through the hidden layers to the output layers and the decision is compared with ground truth labels (Husseini Orabi, et al., 2018).

3.3.6 Neural Networks:

The idea of a neural network is loosely based on the human brain. It contains millions of processing nodes which are densely interconnected. Nowadays most of the nets are organized so as to form a layer of nodes. They are generally feed-forward as mentioned above. Neural networks can also be built using keras module. The neural network consisting of an input, hidden and output layer can be used for classification tasks.

The major advantage of using neural networks is that they do not require much pre-processed data and have a higher computational advantage over traditional machine learning models (Husseini Orabi, et al., 2018). The sequential Keras is a simple yet powerful module to build a neural network which can be used for sentiment analysis, fraud detection and other classification problems.

3.4 Evaluation:

The models that have been implemented will have a set of parameters that define how well the model has performed. Evaluation step takes these parameters into consideration such as precision, accuracy, F1 score and recall. After these parameters have been compared and assessed with the deep learning models implemented, the parameters need to be verified and the findings are summarized. The evaluation parameters usually are baseline parameters with default parameters of the models being used to train the model. Once the baseline metrics are set, the parameters of the models are tuned till the best parameters are identified to train the model. These are known as the hyper-tuned parameters. The effective training time is also a metric that should be considered while training a model.

After considering all the above factors, the findings are evaluated and based on their review, the model that has performed well is recognized. Since the fundamental task is of a classification kind, evaluation parameters such as precision, f1-score, recall and accuracy are considered for this research initiative.

4. Design Specification:

A well-designed and planned architecture is crucial for the implementation of any project. In order to accomplish the goals set out in the beginning of the project, the architecture depicted in Figure 2 has been adopted.

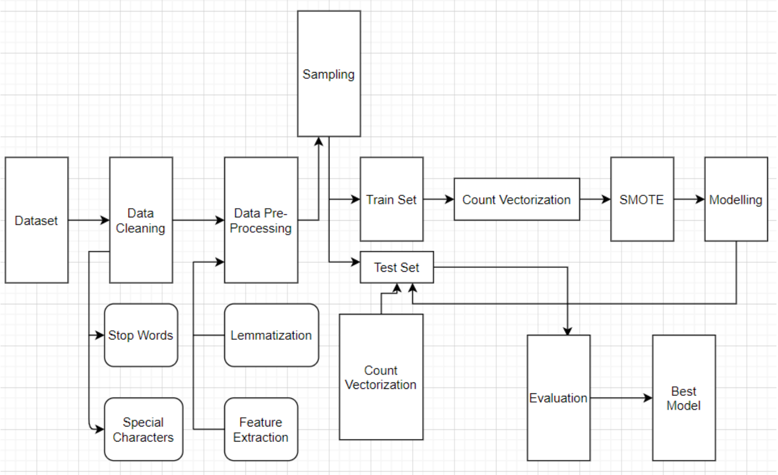

Fig 2. Design Architecture

As seen in the figure, the initial approach begins with the loading of the dataset into the system. As the data is sourced from an online repository, the attributes of the data are well-defined. The features in the dataset need to be cleaned and processed before they can be used in machine-learning model. Cleaning and pre-processing of data is also essential to enhance the performance of machine learning models and also to avoid any chances of overfitting. The data is subjected to a series of cleaning functions which have been developed and also certain features are extracted for further processing. New features are obtained using a combination of the features present. Data preprocessing of textual data consist of methods such as lemmatizing and vectorization of data so that they become interpretable by machine learning models. Random samples of the data have been formed so that the models can be in a time-effective manner. Larger datasets take a huge bulk of computational power. Oversampling techniques have been used to ensure that the target variables are balanced.

After the data has been processed and a sample of 10,000 rows of data have been selected randomly is split into training and testing part in a proportion of 80-20 for the models to be fitted and can learn. This is to establish baseline training and obtain base line evaluation parameters that can later be compared with metrics that have been obtained after parameter tuning of the models. An extensive review of the techniques that have been used for previous researches in mental illness or mental health predictions has been conducted to select models that work well for classification tasks. Models such as SVM, Random Forest, XGBoost , naïve Bayes, MLP classifier and Neural networks have been found suitable to be used considering the nature of the dataset as well as the goal. After models have been trained and performed predictions on the test set, the baseline evaluation parameters can be obtained. The parameters of the models are then hyper-tuned and are trained with these parameters. The optimum parameters are found when the evaluation metrics cannot be further improved. Finally the results that are obtained are documented in forms of visual aids and the most efficient model among them recognized.

5. Implementation:

This section covers a brief explanation of the set of steps that have been carried out for cleaning, pre-processing and transformation of data. The parameters of the models that have been tuned in order to improve performance of the data has also been recorded and their effects explained.

5.1 Data Cleaning:

- Columns such as "Ids", "Query" and "Users" have been dropped. The time feature has been extracted from the date column and stored into a separate feature which will help us classify the mental health. The dataset has been thoroughly searched for neutral sentiments but none were found. The dataset has been checked for null values.
- Since the tweets contain a lot of different characters like numbers, special characters, symbols and tags such as \n and \t, regular expression module has been used to remove them from the text (A. & Sonawane, 2016). After this all the tweets are converted to lower case.
- The next step is to remove stop words that are very common in the English language. The stop words module from the natural language toolkit (Shen, et al., 2017) has been used to remove the 100 most common words from the English language.

5.2 Data Pre-processing:

Lemmatization :

- The tweets were then lemmatized which ensures that the words in the tweets are of the same language and converts it into the root word which stored in the new column labelled "tweets_without_stopwords".

Feature Extraction:

- A new column has been extracted from the time that was separated from the timestamp. The hours after midnight till 0700 i.e 7 in the morning have been identified as critical and the rest of the hours are identified regular. The idea behind classification of mental health is to select those post from the critical hours that have a negative sentiment associated to them. As per (Coppersmith, et al., 2015), since night activity denotes a certain level of mental disturbance a negative comment at this critical period calls for a closer look.

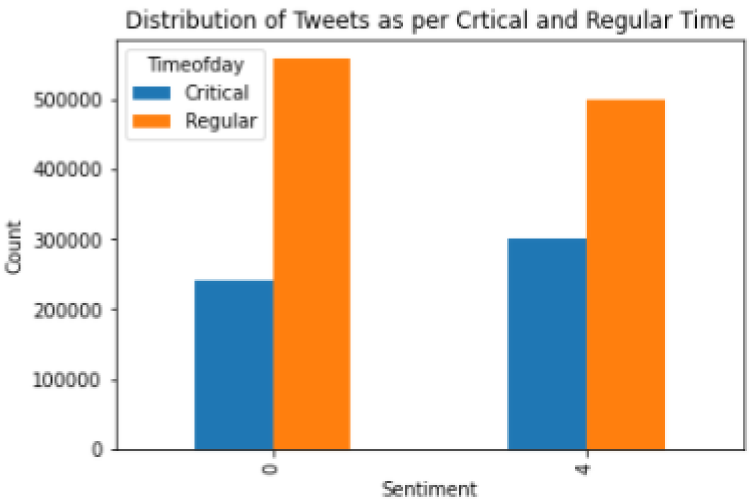

Fig 3. Distribution of tweets during the day and night

Sampling:

- Also, in order to better train the model 10,000 rows have been randomly selected in the first iteration, and then later increased. The samples have been randomly chosen evenly across critical and regular timestamps. A new data frame has been created with the attributes “Target”,”Tweets_without_stopwords”, and “TimeofDay”. This is done to limit the attributes that will be used as input for the model. A new feature “Mental Health” has been created.

- The tweets which have a negative sentiment during critical period has been classified as 1 in the new feature that has been labelled Mental Health and the rest tweets at regular hours have been classified as 0. The data frame has now been split into 8,000 rows of training data and 2,000 rows of test data.

Count-Vectorization:

- The count vectorization module has been used from scikit module to assign or transform the text to a vector. The vectors can be in the form of terms or token counts. The transformation causes the tweets to be transformed into a sparse matrix of numpy elements which are the tokens. This also helps to pre-process the data so that the assigned tokens are used as inputs to the machine learning models.

SMOTE:

- Since only a fraction of the dataset had been chosen the data frame was imbalanced having a larger number of target variable than the other. In order to remove this imbalance, an oversampling technique known as Synthetic Minority Oversampling Technique or SMOTE was used. This ensures that the data is balanced. The data is now optimal for modelling.

5.3 Classification Models:

The models explained in Section 3.3 have been implemented on the training dataset and have been used to perform predictions on test dataset. The first iteration is performed with the default parameters of the models. Then the parameters are hyper tuned in order to improve the performance. A detailed explanation of the parameters tuned with every model has been explained in the following parts.

5.3.1 Support Vector Machine:

Primarily SVM works by creating hyperplanes among datapoints and classifies them accordingly. One of the parameters that has been defined for Support Vector Classifier is the kernel function. The kernel parameter utilizes the already available features and with the help of some transformation creatures some new features. The kernel parameter has been set to "Radial Basis Function" instead of the polynomial function as is suggested in (De Choudhury, et al., 2013).The parameter C is also known as the soft margin cost function, which controls the cost of misclassification when performed on the training data. A higher value of C can potentially overfit the data. Thus, the value of C has been set at 1000 which is an acceptable value. (De Choudhury, 2014). It enables for a soft margin and does not increase the cost of misclassification. The gamma parameter used especially in RBF kernel is responsible to control how fast the decision boundary for datapoints or vectors move implying that higher gamma values mean the decision boundaries that separate the data points are not rigid and have more curvature. The gamma parameter is set to 100. The random state decides the random permutations to split the data points. It provides a seed value for random number generator. All the parameters have been tuned until an optimal performance was achieved.

5.3.2 Naïve Bayes:

The multinomial naïve bayes model has been implemented on the training set using the "MultinomialNB" package from the scikit learn naïve bayes library. After implementing the model on the training data, the model is used to predict on the test set. The classification report function has been used to display the evaluation parameters. As Naïve Bayes has no parameters to be tuned, it can be considered as a baseline model.

5.3.3 Random Forest:

Random forest is an ensemble model which is an upgraded from decision tress. The hyper-parameters that we tune for increasing the efficiency of random forest are mostly the number of decision trees and the number of features which are considered by each tree when splitting a node. 'n-estimators' is the maximum number of trees that you would like to build before taking the averages of predictions. A higher value of estimators will give you a better performance but it will come at a cost of computational power meaning that it will take longer time to train the model. The maximum number of "n_estimators" has been set as 500 since a greater value could marginally improve the performance but the model is then prone to overfitting. Parameters such as "n_estimators",'max_depth' and 'max_features' need to be tuned together as a combination to achieve optimum performance. Increasing the "max_features" as generally increases the performance of the model because at every node there is a higher number of options that can be considered. The downside of this fact is that it reduces the diversity of every individual tree. The max features have been set to auto since it is important to strike a balance between the optimum features. Where "n_estimators" has to compromise between speed and score, 'max_depth' can be used to improve both. In order to prevent overfitting we can limit the depth of the trees. The depth of the trees decide the number of splits it has so that it can capture more information.

5.3.4 XGBoost:

The objective parameter needs to be set to 'binary:logistic' since we need to classify only binary outputs and returns a predicted probability. The seed function is used for producing reproducible results. "n_estimators" defines the number of trees or rounds the model will split into and has been set at 5000. A common issue with tree based boosting algorithms is that they are quick to learn but tend to overfit the training data. An effective way to control the learning is to optimize the learning rate. The value of learning rate should be less that 1. The parameter 'max_depth' is used to limit the maximum depth of the trees. Higher values of "max_depth" are avoided as they influence overfitting. As the depth increases, the tress start learning relations specific to a distinct sample. The value of this parameter is set to 2. Parameters such as "min_child_weight" is the number of samples needed to terminate a branch and a decision node classifying the data. The value of this parameter id set to 50. Since the main goal of a model is to generalize the new data, an optimum "min_child_weight" value reduces the chance of overfitting. A larger "min_child_weight" forces the model to find a path that is common to the entire class, and the default value 1 causes the model to find just a sample common to be classified into the class. The gamma function is used to set the pruning of the tree and also to limit the minimum reduction in loss that can be sustained in a further split of the branch or leaf node. In other words, gamma function is used to decide the complexity of the tree and is set to 10. Subsample ratio decides the part of training data that will be randomly sampled prior to growing trees. If the value is set to 0.5, it implies that half of the training data will be sampled before the model decides to grow trees.

5.3.5 Multi-Layer Perceptron:

MLP classifier is the most basic implementation of a neural network. After the neural network has been used to train the data and it has performed its prediction, the parameters of the classifier are tuned. One of the crucial parameters of a perceptron network is the number of neurons. Changing the hidden layers enables us to set the number of layers of the neurons. A function of (100,100,10) means that the 1st and 2nd layer have 100 nodes each and the 3rd layer has 10 nodes. The "max_iter" function defines the number of times the entire model works through the training dataset. It is also known as the number of epochs defining the time that every sample in the training dataset has updated the internal model parameters. A bigger "max_iter" means a longer training time and to avoid this it has been set to 100. The activation function has been set to rectified linear unit as suggested in (Husseini Orabi, et al., 2018). The learning rate parameter which is set to constant controls the pace at which the model learns a problem. The random state splits the entire training and testing data in a same random order every time the model is run. It can have a significant difference in the output of the model since the datasets are split in random manner with no seed value.

5.3.6 Neural Network:

Learning rate is a crucial parameter which can yield massive improvements without affected the training time. Learning rate should ideally kept below 1 for optimum performance. Larger batch sizes also help improving the model performance as it allows increase in computational speed. However, a massive value of batch size leads to poor generalization and a minimum value could lead to the model learning before analyzing the entire data. The optimizer function has been set to ‘adam’ or Adaptive Moment Estimator. The idea behind it is that the model should not just learn with the minimum criteria with great speed but to decrease the speed for it to have a careful search. It also rectifies the learning rate but is computationally expensive meaning that the resources are used extensively. Since the training data contains a lot of sparse data and is in the form of a csr matrix the loss function has been set to “*sparse_categorical_crossentropy*” which enables the neural networks to work with data containing a lot of zeros. The activation function has been set to ‘relu’ indicating rectified linear units as proposed in (Husseini Orabi, et al., 2018). An epoch is when the entire dataset has been passed through the network back and forth once. Since it is computationally expensive and time consuming to pass the entire data in one epoch they are split in batch sizes. As the number of epochs change based on the dataset, different values of epochs are testes with 10 being the final value.

6. Evaluation:

The sample of 10000 rows which was divided into 8000 rows of training data and 2000 rows of testing data is evaluated based on the performance metrics such as accuracy, precision, F1 score and recall. The evaluation metrics such as accuracy, precision, recall and f1-score have been recorded when the baseline models have been trained and predictions have been performed on the testing data. Then after the parameters are hyper-tuned the values are compared again to see the influence of the parameters on the evaluation metrics.

6.1 Accuracy:

Classification accuracy is often defined as the ration of correct predictions performed by the model to the total number of input samples. Accuracy can also be formulated as:

$$Accuracy = \frac{Number\ of\ \ predictions\ that\ are\ correct}{Total\ number\ of\ predictions}$$ [2]

Equation 1. Accuracy Scores

In classification terms, accuracy is denoted as follows:

$$Accuracy = \frac{TN+TP}{TP+FP+TN+FN}$$ [2]

Equation 2. Accuracy Scores in classification terms

Table 1. Accuracy Scores of models implemented

Model	Baseline Accuracy	Tuned Accuracy
Random Forest	59	79
XGBoost	74	76
MLP Classifier	56	77
Support Vector Machine	63	77
Neural Network	59	77
Naïve Bayes	67	

[2] https://machinelearningmastery.com/how-to-calculate-precision-recall-f1-and-more-for-deep-learning-models

As seen from table 1, the accuracy of the XGBoost model is the highest among all machine learning models. The lowest accuracy is gained by MLP classifier even after the parameter tuning. The most difference in parameter tuning was observed in Support Vector Machine with an upward accuracy of 72% from the baseline accuracy of 63%. Other machine learning models have shown marginal increase of 2% from the baseline accuracy. In terms of the machine learning models, XGBoost can be recognized as the most accurate model. The keras neural network model has faired well in terms of accuracy when compared to its perceptron relative with an accuracy of 59%. There was only an increase of 3% when the parameters were tuned. A lot of factors affect accuracy such as the nature of the dataset and the sampling bias. Accuracy cannot be considered as a role evaluation metric since due to sampling differences the model can predict from samples belonging to the majority class. Hence, it is important to investigate other metrics as well.

6.2 Precision:

This parameter is the measure when the model accurately identifies a class. Our goal is not only to accurately identify the classes where mental health might be compromised but also that the mental health is stable. Below is a table of the comparison between precisions of the classification models applied. Precision can be formulated as below:

$$Precision = \frac{TP}{FP+TP} \quad 2$$

Equation 3. Precision Scores in classification terms

Table 2. Precision Scores of models implemented

Model	Baseline Precision	Tuned Precision
Random Forest	30	31
XGBoost	43	35
MLP Classifier	31	31
Support Vector Machine	30	32
Neural Network	32	33
Naïve Bayes	38	

From table 2, we can infer that random forest and naive bayes algorithms have a higher precision rates in spite of random forest’s accuracy lagging behind the others. XGBoost has the lowest precision rate of 76% and does not increase even after hyper tuning the parameters. Random forest also has the biggest boost in precision after the parameters are hyper tuned. The keras neural network model and the MLP Classifier have the same precision of 77% but when the parameters are tunes the keras model improves it precision by a 3% margin. Random Forest and Naïve Bayes have the best precision scores over all other models.

6.3 Recall:

This parameter is the ratio of true positives upon the total number of true positives and false negatives. In case of this research it corresponds to the ration of cases where the model detected cases of mental illness upon the total number of cases true mental illness and the cases where the person is mentally unstable but was predicted as stable.

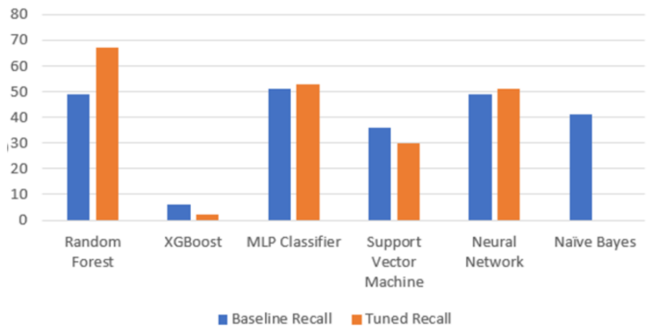

Fig 4. Recall Scores of models implemented

As per the table MLP Classifier has the best recall score which means if it able to identify the true positive cases better. The biggest improvement is also observed in random forest with a 67% recall.

6.4 F1-score:

F1-scores are the harmonic mean of recall and precision. Generally, F1 scores have a lower value than accuracy, as both precision and recall are embedded into the calculations.

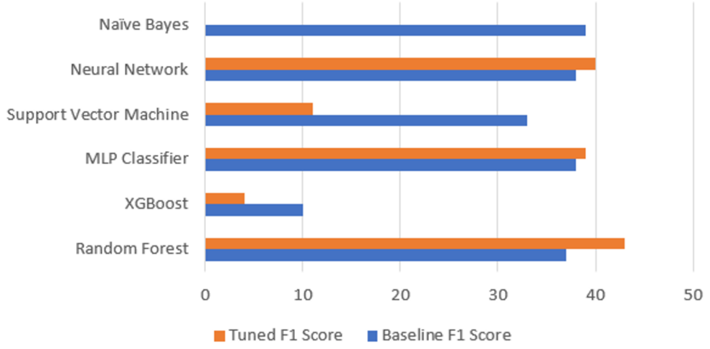

Fig 5. F1 Scores of models implemented

Naïve Bayes has the best F1-score as compared to all other models. However, the most improvement from hyper-parameter tuning has been shown by random forest which achieves a score of 67%.

7.Results and Findings:

The twitter dataset contains a mix of positive and negative comments in different timestamps and the mental health is classified as normal or abnormal based on both these factors. An extensive review of past researches has been carried out to understand the different symptoms of mental illness one of which is lack of sleep. Research into different classification machine learning models have been carried out for prediction of the mental health based on the data available in the dataset. Also, the most common words used by both the classification groups have been extracted and displayed in the figures below. Prediction of mental health along with the identification of the common work makes in a novel work.

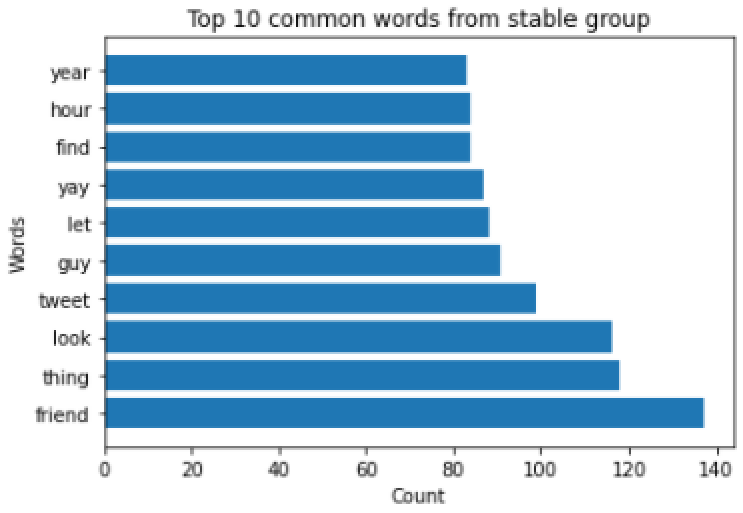

Fig 6. Common words from group classified as stable mental health

A seen from the above figure, this bar chart has been plotted from the 10 most common words used by individuals where mental health was classified as normal. The words are common with day-to-day activities and represent a positive sentiment.

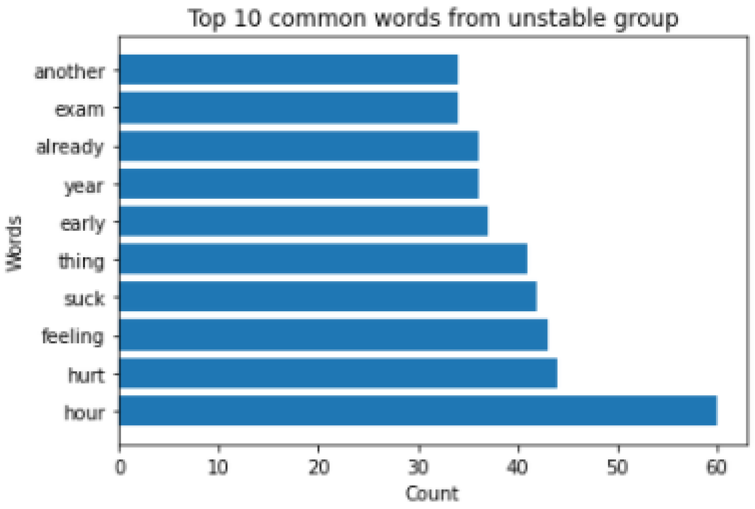

Fig 7. Common words from group classified as unstable mental health

If figure is observed closely, it can be deduced that they contain negative extreme words such as ‘suck’ and ‘hurt’. Also, a word like ‘exam’ could be the source of someone’s stress or anxiety. These words have been extracted from the group where mental health was classified as abnormal. Thus, it can be said with confidence that classification of the groups has been successful.

For classification and prediction, a range of machine learning models along with ensemble techniques have been implemented to identify the most optimum model for this task. Also, keras implementation of a neural network has been carried out to compare the findings with traditional machine learning models. The models have been compared against a set of common evaluation parameters. Precision, recall, f1-score and accuracy are used as standard measures to compare the results and performance of the models. When accuracy scores of machine learning models are compared, XGboost models outputs the highest accuracy in baseline models. However random forest has shown the most improvement when the parameters are tuned. However, accuracy alone cannot determine the performance of a model as it cannot identify overfitting in a model which is why we have also investigated other metrics. As per the precision scores, XGBoost has the lead in precision. Hyper parameter tuning has shown the biggest effect in random forest as it leads the recall score for parameter tuned algorithm. A slight increase of precision score is observed across all algorithms after parameter tuning apart from XGBoost where a drop is observed in precision scores. Also, the recall and precision scores of XGBoost seem to be lagging behind the other machine learning models. A number of reasons are possible for this lag in values. One of the reasons for this is speculated that the model is not able to converge. Another reason for XGBoost giving low values is that gradient boosting requires training the data with a combination of number of iterations and number of classes or trees, while other algorithms such as random only require number of iterations. A major shortcoming of XGBoost is also that it is sensitive to outliers because every classifier is required to rectify the errors made by the predecessor. Also, the method can hardly be scaled as the estimators base the predictions on previous predictions which makes it difficult to streamline. Since only a sample of data has been taken the above issues are faced by the model in basing its predictions.

The recall and f1 scores are also considered in order to recognize the best model for this dataset. The scores for all models have shown improvement after parameter tuning except for SVM. The possible reasons for this are the change of cost functions C and gamma cause the model to underperform. Also, SVM requires feature scaling which means normalizing the data to handle highly varying values. If it is not done the model will weigh large values as higher and smaller values as low irrespective of the units. SVM generally underperforms for larger datasets. It is observed that naïve bayes has the highest scores in recall and f1, but a hyper-tuned random forest has the best performance in terms of recall and f1 score. It can be confirmed from the metrics that MLP Classifier is the best fit for the dataset but a more comprehensive tuning of parameters for a random forest model makes it a much suitable model for this research. A 5-fold cross validation was applied on MLP Classifier and a tuned Random Forest since they have the best performance and it was found that the MLP classifier was performing well on the train data but it was underperforming on the test data. However, in Random Forest, the difference between test score and train score tend to be decreasing after 3rd fold. Thus, we can say that a tuned Random Forest is better at generalization and an optimum fit for the data.

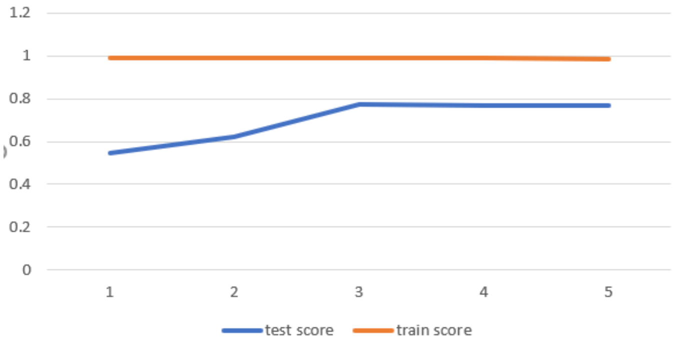

Fig 8. Cross Validation scores of MLP Classifier

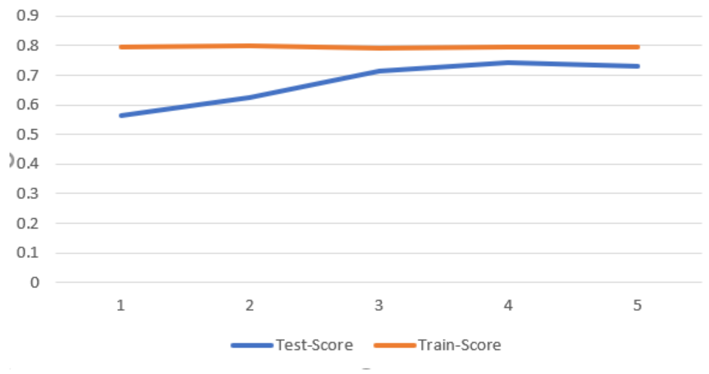

Fig 9. Cross Validation scores of a tuned Random Forest Classifier

Future scope which is discussed in the next section details the techniques that can be employed in order to tackle the limitations present.

8. Discussion

As the main aim of this research has been to classify the Mental health of online users , the models that are employed for this purpose can identify them based on certain characteristics such as sentiment and time of the tweets. The important part that needs to be discussed there is the performance of the various machine learning models applied. Different parameters have been chosen and scrutinized for identifying the ideal model which can predict the mental health of an online user. As per the values obtained in Section 6, the accuracy and precision scores of XGBoost and random forest are decent but XGBoost lags behind in metrics such as precision and recall. Th reason behind this could be due to a number of factors such as the inability of a model to converge, the absence of enough iterations and classes and also due to its susceptibility to outliers. It is also observed that along with random forest, neural network based MLP classifier (Husseini Orabi, et al., 2018) where has performed well across all metrics. The reason for some models underperforming could be attributed to the nature of the dataset. As we are aware that the chosen dataset has minimum attributes that can be used an input for the machine learning models. Since some models excel when a large number of attributes are given for learning as various patterns can then be extracted and used for prediction purposes. Since MLP Classifier and random forest have both depicted the best performance, a 5 fold k-validation is applied to both the algorithms and it was observed that the random forest has a similar score of train and test after the 3rd fold implying that the model predicts just as good on both the training and test data. A detailed outline of the elements that could be considered for further improving the scope of this project has been mentioned in the next section.

9. Conclusions and future work

This research has been carried out with an objective of identifying abnormalities in the mental health of online users using a twitter dataset which contains comments which are positive as well as negative in nature. The classification of mental health revolves around the fact that a negative comment in the critical hours of the day which are identified as midnight and early morning calls for a closer look at the mental health of the user. Various machine learning models have been used to predict the mental health of the users based on their tweets. The data was cleaned and pre-processed as per steps mentioned in (A. & Sonawane, 2016) and the machine learning models were implemented. Models with their default parameters were implemented and their evaluation metrics such as accuracy, precision etc. were recorded. MLP classifier has the best all-round scores in terms of accuracy, F1-score, precision and recall. The models were examined for overfitting using cross validation and an indication of overfitting was observed. The parameters of the models were then hyper tuned to improve their performance and minimize overfitting. Random forest showed the best improvement in terms of all metrics as compared to other models. All models except XGBoost and Support Vector Machine have shown improvement in their performance and the possible reasons for this have been detailed in Section 7. Neural network implementation has yielded results which are virtually the same as compared MLP classifier but it does lag in the section where the parameters are hyper tuned. The noticeable performance of MLP Classifier and Random Forest Classifier and the lagging values in other models can be attributed to a number of different reasons. For instance, the nature of the dataset might be more conducive for classifiers like random forest as explained in section 7. Lack of other features also impact the performance of the models.

Hence, future work could include taking age group, gender and other information about the users to conduct the research. An increase in the number of features has a good chance of improving the model's performance. Also advanced models such as Convolutional Neural Networks can be implemented on a dataset that includes the above listed features. Their performance can be measured with the existing model and the best model can be identified.

9.Acknowledgement

I would like to express my sincere gratitude to my supervisor Prof. Vladimir Milosavljevic for his continued support and guidance throughout the implementation of this research project. His sessions were genuinely informative and the suggestions provided by him have only helped to improve the standard of this research project, His effort to understand the research has always been uplifting and his calm composure has always encouraged me to reach out to him for his inputs.

Also, I would like my parents for their continued support and blessings in this journey of pursuing my Masters in Data Analytics.

References

A., V. & Sonawane, S., 2016. Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications,* 139(11), pp. 5-15.

Aldarwish, M. M. & Ahmad, H. F., 2017. *Predicting Depression Levels Using Social Media Posts.* s.l., s.n.

Bhaugra, D., Till, A. & Sartorius, N., 2013. What is Mental Health. *International Journal of Social Psychiatry,* 59(10.1177/0020764012463315), pp. 3-4.

Coppersmith, G., Dredze, M. & Harman, C., 2015. *Quantifying Mental Health Signals in Twitter.* s.l., Association for Computational Linguistics (ACL).

Corrigan, P. W., Druss, B. G. & Perlick, D. A., 2014. The impact of mental illness stigma on seeking and participating in mental health care. *Psychological Science in the Public Interest,* 15(10.1177/1529100614531398), pp. 37-40.

De Choudhury, M., Counts, S. & Horvitz, E., 2013. *Conference on Human Factors in Computing Systems - Proceedings.* s.l., 10.1145/2470654.2466447.

De Choudhury, M., Counts, S. & Horvitz, E., 2013. *Social media as a measurement tool of depression in populations.* s.l., Association for Computing Machinery.

De Choudhury, M. S. A., 2014. *Characterizing and predicting postpartum depression from shared facebook data.* s.l., Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW.

Guntuku, S. C. et al., 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences,* 18((10.1016/j.cobeha.2017.07.005)), pp. 43-49.

Houghton, D. J. & Joinson, A. N., 2012. *Linguistic markers of secrets and sensitive self-disclosure in Twitter.* s.l., IEEE Computer Society.

Husseini Orabi, A., Buddhitha, P., Husseini Orabi, M. & Inkpen, D., 2018. *Deep Learning for Depression Detection of Twitter Users.* s.l., Association for Computational Linguistics (ACL).

Joseph, A. J. et al., 2015. #Schizophrenia: Use and misuse on Twitter. *Schizophrenia Research,* 165(2-3), pp. 111-115.

Kim, A. E. et al., 2017. Methodological considerations in analyzing twitter data. *ournal of the National Cancer Institute - Monographs,* Issue 47, pp. 140-146.

Lin, L. Y. et al., 2016. ASSOCIATION between SOCIAL MEDIA USE and DEPRESSION among U.S. YOUNG ADULTS. *Depression and Anxiety,* 33(4).

Naslund, J. A., Aschbrenner, K. A., Marsch, L. A. & Bartels, S. J., 2016. The future of mental health care: Peer-To-peer support and social media.. *Epidemiology and Psychiatric Sciences,* 25(2), pp. 113-122.

Prince, M. et al., 2007. No health without mental health. *Lancet, ,* 370(9590), pp. 859-877.

Saravia, E., Chang, C. H., De Lorenzo, R. J. & Chen, Y. S., 2016. *MIDAS: Mental illness detection and analysis via social media..* s.l., Institute of Electrical and Electronics Engineers Inc..

Shen, G. et al., 2017. *Depression detection via harvesting social media: A multimodal dictionary learning solution.* s.l., IJCAI International Joint Conference on Artificial Intelligence.

Spiekermann, S., Acquisti, A., Böhme, R. & Hui, K. L., 2015. The challenges of personal data markets and privacy. *Electronic Markets,* 25(2), pp. 161-167.

Vatrapu, R., Mukkamala, R. R., Hussain, A. & Flesch, B., 2016. Social Set Analysis: A Set Theoretical Approach to Big Data Analytics. *IEEE,* Volume 4, pp. 2542-2571.

Webb, H. et al., 2017. *The ethical challenges of publishing Twitter data for research dissemination.* s.l., WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference.

Westerhof, G. & Keyes, C., 2010. Mental illness and mental health: The two continual model. *Journal of Adult Development,* 17(2), pp. 110-119.