Jadavpur University

A Demographic Study of Depression Analysis in India using Social Media

by

Md Shadab Raza

Examination Roll No. - M4CSE22007
Registration No. - 154131 of 2020-2021
Class Roll No. - 002010502007
Session - 2020-22

This dissertation is submitted for the degree
of Master of Engineering

Under the Guidance and Supervision of

Dr. Sudip Kumar Naskar
Department of Computer Science and Engineering
Jadavpur University
Kolkata - 700 032

June 2022

Declaration of Authorship

I, hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of Master in Computer Science and Engineering studies. All information in this document have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials that are not original to this work.

Signature: _____ Date: _____

Name: Md Shadab Raza

Examination Roll No. - M4CSE22007

Registration No. – 154131 of 2020-2021

Class Roll No. - 002010502007

Session - 2020-22

Thesis Title: A Demographic Study of Depression Analysis in India using Social Media

Certificate of Recommendation

This is to certify that the dissertation entitled “Depression Analysis of Different Cities in India using Social Media” has been carried out by Md Shadab Raza (University Registration No.: 154131 of 2020-2021) under my guidance and supervision and be accepted in partial fulfilment of the requirement for the Degree of Master in Computer Science and Engineering. The research results presented in the thesis have not been included in any other thesis submitted for the award of any degree in any other University or Institute.

Dr. Sudip Kumar Naskar
Thesis Supervisor
Associate Professor
Dept. Of Computer Science and Engineering
Jadavpur University

Signature: _____

Prof. Anupam Sinha
Head of the Department
Dept. Of Computer Science and Engineering
Jadavpur University

Signature: _____

Prof. Chandan Majumdar
Dean
Faculty of Engineering and Technology
Jadavpur University

Signature: _____

Certificate of Approval

This is to certify that the thesis entitled ”Depression Analysis of Different Cities in India using Social Media” is a bonafide record of work carried out by Md Shadab Raza (University Registration No: 1154131 of 2020-2021) in partial fulfillment of the requirements for the award of the degree of Master in Computer Science and Engineering in the Department of Computer Science and Engineering, Jadavpur University, during the period of Sept 2021 to June 2022. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, the opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

External Examiner:

Signature: _____ Date: _____

Dr. Sudip Kumar Naskar
Dept. Of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Signature: _____ Date: _____

Acknowledgement

I would like to express my sincere gratitude to my advisor, Dr. Sudip Kumar Naskar, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University, for his continuous support, patience, motivation, enthusiasm and immense knowledge. His guidance and encouragement helped me in writing the thesis paper. I would also like to thank my batch mates and my family members for their continued encouragement and support.

Md Shadab Raza

Signature: _____

Abstract

Mental illness is one of the most pressing public health issues of our time. The pervasiveness of social media and the near-ubiquity of mobile devices used to access social media networks offer new types of data for understanding the behaviour of people. Depression is typically diagnosed as being present or absent. However, depression severity is believed to be continuously distributed rather than dichotomous. In this work, we focus on applying natural language processing (NLP) techniques to analyse tweets in terms of Depression. In this study we have analyzed 3,44,300 tweets using geolocation feature from 25 different cities in India. We trained deep models that measure depression score of each tweet ranging from 0 to 1. From the tweet-level depression scores, we computed user-level depression scores and city-specific depression index. Furthermore, we collected the census data of India for all these cities and studied the relationship between depression and the socio-economic factors recorded in the census data. The study gives us important insights about the reasons that are causing depression.

Table of Contents

Chapter 1

Introduction

This research sets out to investigate different reasons for depression using twitter data and census data of India. In other words the task is to analyse twitter data of different cities in India and identify what could be the different reason of depression among the people of India.

This chapter is divided into nine sections. In Section 1.1 discusses the motivation and background of the research. In Section 1.2 the problem definition of the research is described and Section 1.3 considers the research aims and objectives. In Section 1.4 the research methodology is outlined. The scope of the research and contributions of the research to knowledge are covered in Sections 1.5 and 1.6 respectively. Lastly, the outline of the thesis is presented in Section 1.7.

1.1 Motivation & Background

Mental illnesses are on the rise all around the world. In the recent decade, there has been a 13 percent increase in mental health and substance use disorders, owing primarily to demographic shifts (to 2017) by World Health Organisation (WHO) [1]. Mental illnesses now account for one out of every five years spent disabled. Around 20% of the world's children and adolescents suffer from mental illness, with suicide being the second largest cause of mortality among those aged 15 to 29. In post-conflict contexts, about one out of every five people suffers from mental illness.

Mental health issues can have a significant impact on all aspects of life, including school or work performance, relationships with family and friends, and community participation. Depression and anxiety, two of the most common mental health illnesses, cost the global economy $1 trillion each year.

Despite these statistics, the global median of government health spending on mental health is less than 2%.

Despite improvements in some nations, people with mental illnesses are frequently subjected to severe human rights breaches, discrimination, and stigma.

1.2 Problem Definition

The outbreak of coronavirus disease 2019 (COVID-19) recently has affected human life to a great extent. Besides direct physical and economic threats, the pandemic also indirectly impact people's mental health conditions, which can be overwhelming but difficult to measure. The problem may come from various reasons such as unemployment status, stay-at-home policy, fear for the virus, and so forth.

After knowing the above problem, we can divide problem into two parts – Finding the extent of depression and finding the different reasons for depression.

1.3 Research aim & objectives

The aims and objectives of the research can be summarized in three parts:

1. Fetching tweets of different users from Twitter to analyse the depression.
2. The design of an algorithm for detecting depression score from tweets.
3. The design for finding correlation between depression and census data so that we can find the reasons for depression.

1.4 Research Methodology

In order to achieve the research objectives, objectivism was adopted as the epistemological stance of the research and positivism as the theoretical perspective. The methodology employed is that of algorithmic approach and the methods are as follows:

Content analysis, i.e. the analysis of related literature and of social media content.

Pilot study, i.e. the prototype of the experiment at the early stage of the research served as a medium to construct dataset for training & testing.

Sampling, data samples were used unbiased from time to time to strengthen the robustness of the algorithm.

Experimentation was undertaken to find the algorithmic approach and the best way of testing the samples.
Statistical analysis was performed on the results in order to compare the efficiency of the algorithms.

1.5 Scope of study

The research will focus exclusively on:
1. Fetching tweets location wise, finding depression score for each tweets and finding average depression score of different cities of India.
2. Collecting census data for different cities and evaluating the relation between depression score and census data and finding reasons for depression.

1.6 Contribution to knowledge

The thesis presents a concept of finding tweets using some keywords related to depression from different geolocation using longitude and latitude of different cities. Developing and algorithm to measure the extent of depression of each tweets and finding average depression score a particular cities. Furthermore gathering data of different cities from census data which is provided by government of India and finding the reasons for depression.

1.7 Thesis Outline

The research carried out in this thesis is explained in four chapters. Chapter 2 reviews the previous related work in this area. Chapter 3 give details about experiments, it contains the methodology of different models, dataset used in this thesis, experimental setup and results and analysis. The thesis concludes in Chapter 4 with a discussion of the contribution made by this research as well as recommendations for future work.

Chapter 2

Literature Survey

There have been significant amount of research on Mental Health Analysis of Social Media. The most relevant ones are studied in this chapter. For a structured survey, we divided these works into three categories - Detecting Mental Depression from Social Media (Section 2.1), Mental Depression and Socio-Economic Factors (Section 2.2), and Social Media Addiction & Mental Depression (Section 2.3).

2.1 Detecting Mental Depression from Social Media

Glen Coppersmith et al. [2] present a novel study on mental health phenomena in publicly available Twitter data, demonstrating how rigorous application of simple natural language processing methods can yield insight into specific disorders as well as mental health in general, as well as evidence that yet-to-be-discovered linguistic signals relevant to mental health exist in social media. We offer a unique way for quickly and affordably collecting data for a variety of mental disorders, then focus on four in particular: post-traumatic stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD). All of the information they obtain is public, having been posted between 2008 and 2013, and has been made available through Twitter's application programming interface (API). They provide three types of experiments. First, they evaluate their data collection approach by utilising LIWC to replicate past findings. They then compare and develop classifiers to identify each group from the control group, proving that there is useful signal in each group's language. Finally, they examine the linkages between their analytics and classifiers in order to extract insight into quantifiable and meaningful mental health signs in Twitter.

Irene Li et al. [3] use natural language processing (NLP) tools to assess tweets in terms of mental health in this study. Anger, anticipation, disgust, fear, joy, sadness, surprise, and trust are the emotions that deep models classify each tweet into. They created the EmoCT (Emotion-Covid19-Tweet) dataset by manually classifying 1,000 English tweets for training purposes. They also

suggest a method for determining the causes of sadness and fear, as well as studying the emotion trend at both the keyword and topic level. They used Twitter API to run a crawler with a list of keywords such as coronavirus, covid19, covid, COVID-19, covid 19, etc. To get language and geographical statistics, they looked at 8,148,202 tweets from March 24 to March 26, 2020. They performed single-label and multi-label classification tasks using this dataset, with promising results. They used two approaches to determine keyword correlations and conducted some analysis to investigate the emotion trend to better understand why people could be sad or fear.

Depression, rather than being an illness that one has or does not have, might be understood as a continuous construct that changes through time. H. Andrew Schwartz et al. [4] developed a regression model based on survey responses and status updates from 28,749 Facebook users that predicts people' degree of depression based on their Facebook status updates. Their user-level predictive accuracy is modest, but it outperforms an average user sentiment baseline significantly. They apply their model to estimate user changes in depression across time, and they find that, in line with the literature, users' depression levels tend to rise from summer to winter. They show the possibility to examine factors driving people' level of depression on a monthly or even weekly basis by looking at its most highly connected language features with the further development of regression models.

Munmun De Choudhury et al. [5] present how to use crowdsourcing to create a massive corpus of tweets shared by people who have been diagnosed with clinical depression. Then, using this corpus as a training set, they create a probabilistic model to see if posts indicate depression. The model is based on signals from Twitter that show social activity, emotion, and language. They develop a social media depression index based on the model, which might be used to assess depression levels in communities. The measure's geographic, demographic, and seasonal patterns of depression confirm psychiatric findings and are highly correlated with depression statistics published by the Centers for Disease Control and Prevention (CDC). A total of 489 users were obtained, all of whom claimed that they either had clinical depression that began in or after September 2011 and ended before June 2012, or that they had never had depression before. There were 251 males and 238 females in the group, with a median age of 25.

Sharath Chandra Guntuku et al. [6] shows recent research that used social media to predict mental illness are discussed. Screening surveys, public publication of a diagnosis on Twitter, or membership in an online forum were used to identify mentally ill users, who were distinguished from control users by patterns in their language and online activities. Through large-scale passive monitoring of social media, automated detection approaches may help identify depressed or otherwise at-risk individuals. Data ownership and protection, as well as clinical and operational considerations about integration into care systems, should all be addressed as soon as possible.

Ariel Shensa et al. [7] wanted to see whether there were any specific patterns of social media usage (SMU) and if there were any correlations between those patterns and depression and anxiety symptoms. A nationally representative sample of 1730 US individuals aged 19 to 32 took an online survey in October 2014. SMU patterns were discovered using cluster analysis.
In a large sample of US young adults, cluster analysis revealed five unique SMU patterns. In respect of key SMU characteristics, socio-demographic variables, and relationships with depression and anxiety symptoms, the groups differed. Two distinct patterns of use have been linked to an increased risk of depression and anxiety symptoms. Three different patterns of moderate use were not linked to increased sadness or anxiety symptoms in any of the three groups. Individual features, rather than patterns of use, may be more representative of real-world SMU, and hence useful in clarifying correlations between SMU and depression and anxiety symptom levels.

Depression is regarded as the leading cause of global disability and a leading cause of suicide. They examine at Reddit postings to see whether there are any factors that reveal relevant online users' depressed attitudes. To train the data and evaluate the efficiency of their suggested solution, they use Natural Language Processing (NLP) techniques and machine learning algorithms. Michael M. Tadesse et al. [8] discovered a lexicon of terms that are more commonly used in depressed stories. There are depression-indicative posts (1293) and standard posts (548) in the data corpus. Depression-indicative postings are gathered from relatively large subreddits devoted to depression, where depressed persons seek online support. Non-depressed users' standard posts are gathered from subreddits relating to a family or friends. Bigram with the Support Vector Machine (SVM) classifier is the best single feature for detecting depression with an accuracy of 80% and an F1 score 0.80. The Multilayer Perceptron (MLP) classifier best demonstrates the strength and

usefulness of the combined features (LIWC+LDA+bigram), achieving the best performance for depression identification with 91% and 0.93 F1 score.

2.2 Mental Depression and Socio-Economic Factors

Flora I. Matheson et al. [9] shows that residents of "stressed" neighbourhoods have higher levels of depression than residents of less "stressed" neighbourhoods, according to multilevel study. Individual data comes from two cycles of the Canadian Community Health Survey, a national probability sample of 56,428 persons living in 25 Census Metropolitan Areas across Canada, with linked 2001 census of Canada tract information. The Centre for Epidemiologic Studies-Depression Scale Short Form is used to assess depression, with a cut-off of four or more symptoms. Two variables of neighbourhood chronic stress—residential mobility and material deprivation—as well as two indicators of population structure—ethnic diversity and dependency—were found using factor analysis of census tract statistics. A significant contextual effect of neighbourhood chronic stress exists after adjusting for individual-level gender, age, education, marital and visible minority status, as well as neighbourhood-level ethnic diversity and dependency. Depression is linked to the daily stress of living in a neighbourhood where residential mobility and material deprivation are prevalent. They looked at the possibility that women are more reactive to chronic stressors, resulting in a higher risk of depression, because gender frames access to personal and social resources. They did not, however, identify any gender-based variance in depression between neighbourhoods.

Sandeep Grover et al. [10] shows review of depression studies in children and adolescents, depression is a prevalent mental illness that affects people of all ages, including children and teenagers. In children and adolescents, depression is frequently linked to significant disability. According to available data, the point prevalence of depression/affective disorders in clinic-based studies ranges from 1.2% to 21%; 3% to 68% in school-based studies; and 0.1% to 6.94% in community studies. Only one incidence research was conducted in India, and the incidence was reported to be 1.6%.Various studies have reported various education-related challenges, relationship troubles with parents or at home, family-related concerns, economic difficulties, and

other aspects as risk factors for depression. Depressed mood, decreased interest in play activities, concentration difficulties, behaviour problems in the form of anger and aggression, pessimism, decreased appetite, decreased sleep, anhedonia, and somatic symptoms are among the most commonly reported symptoms, according to a small number of studies. None of the Indian research have looked at the efficacy/effectiveness of various antidepressants in depressed children and adolescents.

Manju Pilania et al. [11] set the goal of this systematic review and meta-analysis was to determine the prevalence of depression among India's older population. The articles in this study were published between 1997 and 2016. Studies conducted in special demographic groups, such as hospitals, were eliminated since they only reported a subtype of depression and did not specify the screening tool. Data was extracted from published reports, and authors were contacted for any missing information. Fifty-one research from 16 Indian states were compiled into 56 datasets, with the prevalence of depression among India's older population estimated to be 34.4 %. About one third of India's old population was depressed, with a female preponderance. Estimates differed depending on the type of study tool, geographic region, sample methodology, and whether or not dementia was present. Because the studies included in this evaluation used different methodological approaches and screening technologies, the pooled estimate should be evaluated with caution.

Low socio-economic status is associated to a higher frequency of depression in cross-sectional studies, it is unclear if changes in socio-economic status contribute to changes in depression rates (Lorant et al. [12]). Since most longitudinal studies have been of short duration and have characterised socio-economic status using relatively time-invariant variables such as education or occupational social class (Lynch et al. [13]; Weich & Lewis [14]), to assess whether longitudinal change in socio-economic factors affects change of depression level Vincent Lorant et al. [15] present a prospective cohort study using the annual Belgian Household Panel Survey (1992-1999), depression was assessed using the Global Depression Scale. Material standards of living, education, employment status, and social interactions were all considered socio-economic considerations. Between annual waves, a drop in material standards of living was linked to an increase in depressed symptoms and the occurrence of serious depression. Depression was influenced by a variety of factors. The negative impacts were stronger than the favourable ones;

ceasing to cohabit with a spouse increased depressive symptoms and caseness, whereas improving circumstances reduced them; the negative effects were stronger than the positive ones.

Depression is a significant public health issue which transcends communities and countries. It is the leading cause of disability worldwide, and the global burden of depression is on the rise (Jean-Pierre Lépine er al. [16]). The prevalence of depression varies considerably both within and between countries across Europe [17,18]. Low socio-economic status (SES) has been linked to an increased prevalence of depression. Aislinne Freeman et al. [19] used standardised methodologies and assessments, as well as a composite score for SES, to assess the relationship between SES and depression in three European countries that represent distinct regions across Europe. For Finland, Poland, and Spain, the risks of depression were significantly reduced for every unit rise in the SES index after adjusting for confounders. Furthermore, in each country, better education considerably reduced the risk of depression, but not wealth.

2.3 Social Media Addiction & Mental Depression

Rahmatullah Haand et al. [20] propose a study that looked into the relationship between public media addiction and depression among university students in Afghanistan's Khost region. A 46-item self-administered questionnaire was delivered to 384 students from three universities: Shaikh Zayed, Ahmad Shah Abdali, and Pamir University, using stratified random sampling. The Internet Addiction Test (IAT) by Kimberly Young was used to assess social media addiction, while the Centre for Epidemiologic Studies Depression Scale was used to assess depression (CES-D). The relationship between social media addiction and depression was studied using the Pearson correlation coefficient, simple linear regression, and component analysis. Social media addiction has a positive correlation with depression, and depression strongly predicts social media addiction, according to the data. In both developed and developing countries, the addictive use of social media is positively associated with depression.

While online social media has become inextricably linked to our daily lives, it is being blamed for an increase in mental health concerns among young people. The evidence on the impact of social media use on depression, anxiety, and psychological distress in adolescents was analysed by Betul Keles et al. [21]. It is estimated that 50% of all mental disorders are established by the age of 14

and 75% by the age of 18 (Kessler et al. [22]; Kim-Cohen et al. [23]). According to the Pew Research Centre (Amanda Lenhart et al. [24]), at least 92% of teenagers are active on social media. The results were categorised into four social media domains: time spent, activity, investment, and addiction. All these four categories were found as correlated with depression, anxiety and psychological distress, with an acknowledgement for the complexity of these relationships. Although studies have looked into mediating and moderating elements that may contribute to or exacerbate the proposed relationship, there are still a number of mediators and moderators that have yet to be discovered that could explain the relationship's direction.

Serdar Aydin et al. [25] aimed to investigate the effects of social media addiction on depression in adult individuals. The researchers wanted to see if social media dependence had different effects depending on different variables (age, gender, the highest level of education, duration of daily use of social media, frequency of social media use, etc.). In a study conducted by Kirik et al. [26] on social media addiction with 271 undergraduates, no significant difference was found in terms of gender either. There are also studies in the literature showing that social media addiction of men is higher than in women [27–29]. Uncontrolled internet use may adversely affect the individual's physical, mental, social, and cognitive development [30]. Social media had a large place in their daily lives and negatively affected their lives. In characteristics including the number of children, age, and income, there were significant variations between depression and social media dependency. When social media addiction was analysed in terms of gender among socio-demographic characteristics as a consequence of the study, no significant difference was identified.

The direct and indirect impacts of self-esteem, daily internet use, and social media addiction on adolescent depression levels were explored using a model by Kircaburun et al. [31]. This descriptive study included 1130 students ages 12 to 18 who were enrolled in several schools in the Aegean's southern region. Depression was also linked to a negative relationship with self-esteem and a positive relationship with daily internet use. Adolescent depression was indirectly influenced by social media addiction (positively). This finding coincides with some studies (Aydm & San [32]; Bahrainian et al. [33]; Kim et al. [34]; Mei et al. [35]; Zhang [36]), as well as it contradicts the other ones (Ayas & Horzum [37]; Reisoğlu, Gedik & Göktaş [38]).

Over the last several years, there has been a substantial increase in study into social media addiction, with the majority of studies focusing on "Facebook addiction" (FA), which has been classified as a probable behavioural addiction by some experts (Hormes [39]). Griffiths [40], on the other hand, stated that people's activities on Facebook might include things like gaming and gambling, in addition to social networking. FA is a non-chemical (i.e., behavioural) addiction characterised by excessive human-machine contact (Cerniglia et al. [41]) and the presence of six basic addiction criteria: salience, mood alteration, tolerance, withdrawal symptoms, conflict, and relapse (Griffiths [42-44]). Being single, having less involvement in physical activities, sleep disturbance (more or less than 6 to 7 hours of sleep), time spent on Facebook (more than 5 hours per day), and depression symptoms were all found to increase the chance of being addicted to Facebook in a regression analysis.

Chapter 3

Experiments and Results

The experiment carried out in this thesis is explained in four sections. Section 1 reviews basic definition of algorithm of all the models carried out during this thesis. Section 2 explain all the details about dataset used in this thesis. Section 3 shows experimental setup of all the models. Section 4 contains the results and analysis of all the models.

3.1 Models

3.1.1 Logistic Regression

This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\text{Logistic regression} = \frac{1}{1+e^{-x}}$$

Logistic regression uses a loss function referred to as "maximum likelihood estimation (MLE)" which is a conditional probability. If the probability is greater than 0.5, the predictions will be classified as class 0. Otherwise, class 1 will be assigned.

3.1.2 Decision Tree

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

So among all of the attributes which one should we pick first? The attribute which classifies the training data best will be picked first.

Entropy: Entropy, also called as Shannon Entropy is denoted by H(S) for a finite set S, is the measure of the amount of uncertainty or randomness in data.

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

Intuitively, it tells us about the predictability of a certain event.
In particular, lower values imply less uncertainty while higher values imply high uncertainty.

Information Gain: Information gain is also called as Kullback-Leibler divergence denoted by IG(S,A) for a set S is the effective change in entropy after deciding on a particular attribute A. It measures the relative change in entropy with respect to the independent variables.

$$IG(S, A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

where IG(S, A) is the information gain by applying feature A. H(S) is the Entropy of the entire set, while the second term calculates the Entropy after applying the feature A, where P(x) is the probability of event x.
Constructing a decision tree is all about finding the attribute that has the highest information gain.

3.1.3 Random Forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision.

3.1.4 Multinomial Naive Bayes

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance. The Naive Bayes classifier is made up of a number of algorithms that all have one thing in common: each feature being classed is unrelated to any other feature. A feature's existence or absence has no bearing on the inclusion or exclusion of another feature. It's based on the formula below:

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

3.1.5 Linear SVM

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.
In fact, we have an infinite lines that can separate these two classes. So how does SVM find the ideal one?
According to the SVM algorithm we find the points closest to the line from both the classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. Our goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane.

3.1.6 SVM using RBF Kernel

When the data set is linearly inseparable or in other words, the data set is non-linear, it is recommended to use kernel functions such as RBF.

Kernel Function is used to transform n-dimensional input to m-dimensional input, where m is much higher than n then find the dot product in higher dimensional efficiently. The main idea to use kernel is: A linear classifier or regression curve in higher dimensions becomes a Non-linear classifier or regression curve in lower dimensions.

Mathematical Definition of Radial Basis Kernel:

$$K(x, x') = \exp\left(-\frac{\left||x - x'|\right|^2}{2\sigma^2}\right)$$

where x, x' are vector point in any fixed dimensional space.

But if we expand the above exponential expression, It will go upto infinite power of x and x', as expansion of e^x contains infinite terms upto infinite power of x hence it involves terms upto infinite powers in infinite dimension.

3.1.7 k-Nearest Neighbor

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another. For classification problems, a class label is assigned on the basis of a majority vote—i.e. the label that is most frequently represented around a given data point is used. In order to determine which data points are closest to a given query point, the distance between the query point and the other data points will need to be calculated. These distance metrics help to form decision boundaries, which partitions query points into different regions: Euclidean distance, Manhattan distance, Hamming distance.

3.1.8 XGBoost

XGBoost, which stands for Extreme Gradient Boosting, is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then

fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

3.1.9 Artificial Neural Network (ANN)

ANN are multi-layer fully-connected neural nets that look like the Figure 3.1 below. They consist of an input layer, multiple hidden layers, and an output layer. Every node in one layer is connected to every other node in the next layer. We make the network deeper by increasing the number of hidden layers.

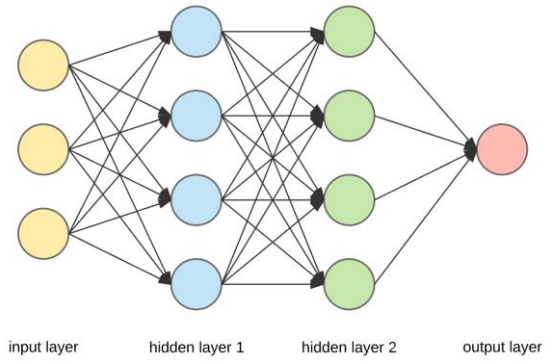

Figure 3.1: Basics of Artificial Neural Network

If we zoom in to one of the hidden or output nodes, what we will encounter shown in the Figure 3.2 below.

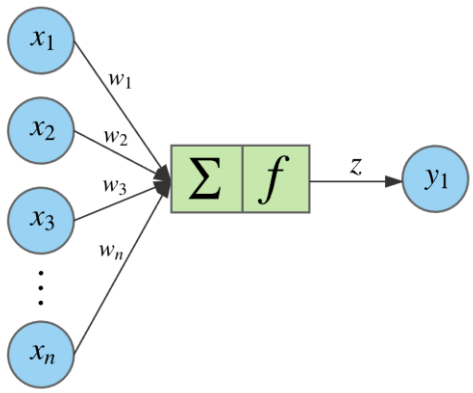

Figure 3.2: Fundamental architecture of ANN

A given node takes the weighted sum of its inputs, and passes it through a non-linear activation function. This is the output of the node, which then becomes the input of another node in the next layer. The signal flows from left to right, and the final output is calculated by performing this procedure for all the nodes. Training this deep neural network means learning the weights associated with all the edges.

3.1.10 Convolutional Neural Network (CNN)

Convolutional neural networks are distinguished from other neural networks by their superior performance with image, speech, or audio signal inputs. They have three main types of layers, which are:

- Convolutional layer
- Pooling layer
- Fully-connected (FC) layer

Convolutional Layer

The convolutional layer is the core building block of a CNN, and it is where the majority of computation occurs. It requires a few components, which are input data, a filter, and a feature map.

Pooling Layer

Pooling layers, also known as downsampling, conducts dimensionality reduction, reducing the number of parameters in the input.

Fully-Connected Layer

This layer performs the task of classification based on the features extracted through the previous layers and their different filters.

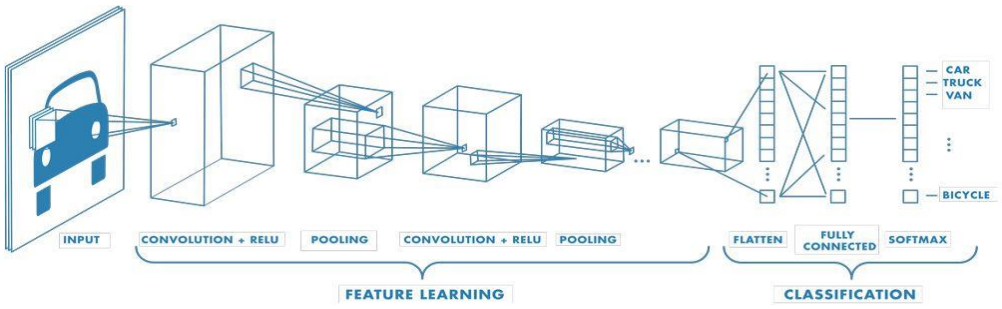

Figure 3.3: Basics of Convolutional Neural Network

3.1.11 Simple Recurrent Neural Network (RNN)

RNNs are a type of neural network that can be used to model sequence data. All of the inputs and outputs in standard neural networks are independent of one another, however in some circumstances, such as when predicting the next word of a phrase, the prior words are necessary, and so the previous words must be remembered. As a result, RNN was created, which used a Hidden Layer to overcome the problem. The most important component of RNN is the Hidden state, which remembers specific information about a sequence.

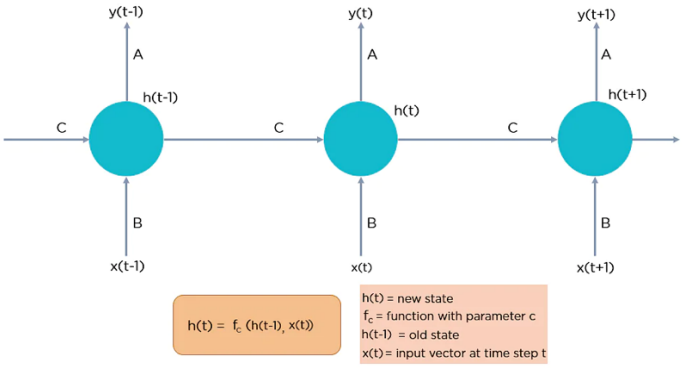

Figure 3.4: Basics of Recurrent Neural Network

Here, “x” is the input layer, “h” is the hidden layer, and “y” is the output layer. A, B, and C are the network parameters used to improve the output of the model. At any given time t, the current input is a combination of input at x(t) and x(t-1). The output at any given time is fetched back to the network to improve on the output.

3.1.12 Long short-term memory (LSTM)

Long short-term memory (LSTM) is a recurrent neural network (RNN) architecture (an artificial neural network) proposed in 1997 by Sepp Hochreiter and Jurgen Schmidhuber [45]. Like most RNNs, a LSTM network is universal in the sense that given enough network units it can compute anything a conventional computer can compute, provided it has the proper weight matrix, which may be viewed as its program. Unlike traditional RNNs, an LSTM network is well-suited to learn from experience to classify, process and predict time series when there are time lags of unknown size and bound between important events. Relative insensitivity to gap length gives an advantage to LSTM over alternative RNNs and hidden Markov models and other sequence learning methods in numerous applications.

3.1.13 Gated Recurrent Unit (GRU)

GRU is an advancement of the standard RNN. It was introduced by Kyunghyun Cho et al. [46] in the year 2014. GRUs are very similar to Long Short Term Memory(LSTM). Just like LSTM, GRU uses gates to control the flow of information. They are relatively new as compared to LSTM. This is the reason they offer some improvement over LSTM and have simpler architecture.
Another Interesting thing about GRU is that, unlike LSTM, it does not have a separate cell state. It only has a hidden state. Due to the simpler architecture, GRUs are faster to train.

3.1.14 Bidirectional LSTM

Bidirectional long-short term memory(bi-lstm) is the process of making any neural network o have the sequence information in both directions backwards (future to past) or forward(past to future).
In bidirectional, our input flows in two directions, making a bi-lstm different from the regular LSTM. With the regular LSTM, we can make input flow in one direction, either backwards or forward. However, in bi-directional, we can make the input flow in both directions to preserve the future and the past information.

3.1.15 Bidirectional Encoder Representations from Transformers (BERT)

BERT was proposed by Devlin et al. [47] in 2018. It is basically a Transformer encoder stack trained on two language modeling tasks – Masked Language Model(MLM) and Two Sentence Task.

1. **Masked Language Model** : A fraction of the input tokens are masked randomly and BERT is trained to predict the masked word. Sometimes, instead of masking a token is replaced with some other token and BERT is trained to predict the correct word.
2. **Two Sentence Task** : This task involves pre-training BERT to decide given two sentences, if the second sentence is likely to follow the first.

BERT has two model sizes (Figure 3.5[1]) —

---

[1]https://jalammar.github.io/illustrated-bert/

• **BERT Base** - contains 12 encoder layers and 12 attention heads.

• **BERT Large** - contains 24 encoder layers and 16 attention heads.

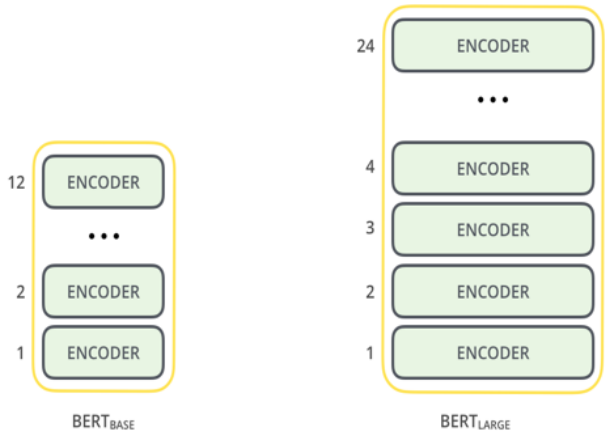

Figure 3.5: The two model sizes of BERT

Each encoder is also known as Transformer block and consists of a self-attention layer and a feed-forward neural network. An encoder receives a list of vectors which is passed into the self-attention layer followed by the feed-forward network. The output produced by one encoder is then forwarded to the next. For the first encoder, the input is the embedding vector added with the positional encoding for the input.

Self-attention layer. The self-attention layer creates an encoding for each token by incorporating information from other positions in the input sequence. The output of this layer is generated using a series of calculations explained below.

- For each word, three vectors are created – query vector, key vector and value vector.

  $q_i=x_i\times W_Q$, $k_i=x_i\times W_K$, $v_i=x_i\times W_V$

  where x_i is the embedding vector for the i-th word.

- Then for each word, we calculate a score against every other word in the input. This is given by the dot product of the query vector and the key vector of the word we are scoring.

$$score_{ij} = q_i \cdot k_j$$

Here, $score_{ij}$ is the score for the j-th word w.r.t. the i-th word.

- The scores are then passed through a softmax function. These values determine the relevance of each word with respect to a particular word.
- Each softmax score is then multiplied with the respective value vectors. The products are then summed up to obtain the self-attention output for the particular position.

BERT uses multiple attention heads which implies that each head maintains separate query, key and value weight matrices and therefore produces multiple encoding z for one input word. These encodings are multiplied with another weight matrix W_O to obtain the final output.

$$z = Z \cdot W_O$$

where Z is the concatenated output from all the attention heads. The output z is then passed through a normalisation layer and then fed into the feed-forward neural network. Figure 3.6[2] gives a visual representation of an encoder.

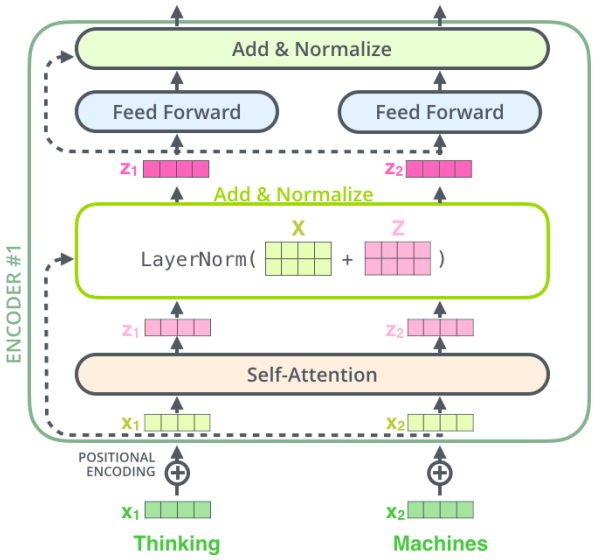

Figure 3.6: A single encoder block

[2]http://jalammar.github.io/illustrated-transformer/

3.1.16 Correlation Technique

After using above models we get depression score of different cities then we need to analyse the census data and relationship between depression and census data so that we can able to find the reasons behind the depression.

The statistical relationship between two variables is referred to as their correlation. A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. Correlation can also be neutral or zero, meaning that the variables are unrelated.

3.1.16.1 Pearson's Correlation

The Pearson correlation coefficient (named for Karl Pearson) can be used to summarize the strength of the linear relationship between two data samples.

The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample. It is the normalization of the covariance between the two variables to give an interpretable score.

Pearson's correlation coefficient = covariance(X, Y) / (stdv(X) * stdv(Y))

3.1.16.2 Spearman's Correlation

Two variables may be related by a nonlinear relationship, such that the relationship is stronger or weaker across the distribution of the variables.

Spearman's correlation coefficient = covariance(rank(X), rank(Y)) / (stdv(rank(X)) * stdv(rank(Y)))

3.1.17 Principal component analysis (PCA)

As we know in PCA principal components are perpendicular to each other. Hence we can say that if the PC1 has maximum loading from let say feature1 and featue2 and there features has very less loadings in other PC then we can say that these two feature are really close in feature space that's

why they are contributing much in formation of PC1 or in other words they are somehow correlated as they are close in feature space.

3.2 Datasets

We performed our experiment for depression analysis on social media dataset and census data. We have used 3 datasets: First, we use labelled dataset of tweets for training the model. Second, we use unlabelled random tweets from different cities in India. Last, we use the 2011 Census of India for finding the reasons for depression.

3.2.1 Dataset 1

We use the dataset from kaggle website which was created by Creative Commons which is a non-profit organization that helps overcome legal obstacles to the sharing of knowledge and creativity to address the world's pressing challenges. The dataset contains 18047 tweets, each containing one or more sentences are labelled as '0' for not depressed and '1' for depressed.
The dataset is provided in csv file format and each tweets has a unique id.

3.2.2 Dataset 2

There are two essential kind of dataset that are required in this dataset: first, we identify people related to depression by the web scraper using the keyword #depression, #ptsd, #mentalhealth and related keywords by scraping all tweets and second, after getting user_id of the person we have analyse 100 most recent tweets for each users which we get from previous data so that we can get more information about that person.

All the tweet used are public and we get all the details of users from twitter APIs and all the tweets were fetched in a span of 1 week from 23rd April, 2022 to 29th April, 2022. We have scrap tweets for 25 different cities of India using longitude and latitude of the city which is a provided feature in Twitter API. Dataset contains 3443 unique users and around 3,44,300 tweets were collected for this experiment.

3.2.3 Dataset 3

We use 2011 census of India data of 25 cities in India and analyse all possible features that can cause depression. We use 82 features from census data like literacy rate, sex ratio, number of Non-Marginal workers, etc. and depression score for all the cities and create a csv file dataset.

3.3 Experimental Setup

In this section we show all the details about hyperparameter of the models. Data pre-processing is an important aspects for achieving good accuracy, vector size, number of neurons in the network, number of epoch used all these are important for better accuracy.

3.3.1 Tools Used

3.3.1.1 Twitter API v2

The Twitter API enables programmatic access to Twitter in unique and advanced ways. We will use the Twitter API to access data about both Twitter users and what they are tweeting about. Tweepy is an easy-to-use python library for accessing the twitter API.
Tweepy provides several different methods to refine our queries like send a tweet, search tweets, scraping with advanced queries like search user or tweets by geographic location.

3.3.1.2 Scikit-learn

Scikit-learn is a free software machine learning library for the Python programming language. Simple and efficient tools for predictive data analysis and built on NumPy, SciPy, and matplotlib. We use for data preprocessing like feature extraction, normalisation and we use for machine learning algorithm for classification and improving parameter via parameter tuning and comparing models with metric function.

3.3.1.3 TensorFlow

TensorFlow is a deep learning library. TensorFlow allows us to perform specific ma- chine learning number-crunching operations like derivatives on huge matrices with large efficiency. We

can also easily distribute this processing across our CPU cores, GPU cores, or even multiple devices like multiple GPUs. We can even distribute computations across a distributed network of computers with TensorFlow. So, while TensorFlow is mainly being used with machine learning right now, it actually stands to have uses in other fields, since really it is just a massive array manipulation library.

3.3.2 Data Pre-processing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. There are two parts before applying models on data, first is cleaning the data and second is convert text data to integer.

3.3.2.1 Data Cleaning

After reading data, first thing is to remove unnecessary features in the dataset. After that we need to remove @ sign because many text contains @username but for analysing depression this is not important to consider. Another thing in tweets which is redundant is URL and email ids, so we need to remove that from the text and also removing special character from the text.

Next step is tokenization, we split the sentence into group of words and then we use Porter stemming algorithm for removing the commoner morphological and inflexional endings from words in English.

3.3.2.2 Text into Numeric Encoding

Text is a sequential data and we need to represent each word as numeric data and aggregate into vector. We use term frequency-inverse document frequency (TF-IDF) for converting text to numeric vector for machine learning models and Word Embeddings which uses neural networks like word2vec for deep learning models because it captures the relationship between the words and they are called feature vectors and then they were feed into neural networks.

3.3.3 Model Implementation

After data pre-processing and text encoding we need to split data into training set and testing set. For splitting we use stratified k-fold cross validation techniques where we put value of k is from 2 to 10. After splitting we tune parameter for different models and we will see parameter tuning in this section.

3.3.3.1 Logistic Regression

We have implemented logistic regression after pre-processing the data using sklearn library in python language. We use L2 Regularization for penalty to addressing over-fitting and value of C is 30 which is inverse of regularization strength, smaller values specify strong regularization. We choose solver algorithm to optimization is 'lbfgs' and maximum number of iterations taken for the solvers to converge is 1000.

3.3.3.2 Decision Tree

We use the gini index to measure the quality of split and 'best' splitter strategy used to choose the split at each node. The maximum depth of the tree is 'None', then nodes are expanded until all leaves are pure or until all leaves contain less than min samples split which is 2 and the minimum number of samples required to be at a leaf node is 1.

3.3.3.3 Random Forest

We use number of trees in the forest is 100 and gini index to measure the quality of split. The maximum depth of the tree is 'None', then nodes are expanded until all leaves are pure or until all leaves contain less than min samples split which is 2 and the minimum number of samples required to be at a leaf node is 1. Bootstrap samples is are used when building trees.

3.3.3.4 Multinomial Naive Bayes

The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work. We use smoothing parameter alpha is 1.0 and learning class prior probabilities is true.

3.3.3.5 Linear SVM

We use regularization parameter C as 1.7 and the penalty is a squared l2 penalty. Kernel used in this model is linear and value of gamma is ‘scale’.

3.3.3.6 SVM with RBF Kernel

We use regularization parameter C as 1.75 and the penalty is a squared l2 penalty. Kernel used in this model is radial basis kernel (rbf) and value of gamma is ‘scale’.

3.3.3.7 K-Nearest Neighbor

Number of Neighbor used int this model is 11 and weight parameter is uniform i.e., all points in each neighborhood are weighted equally. ‘auto’ will attempt to decide the most appropriate algorithm based on the values passed to fit method. The distance metric to use for the tree is minkowski, and with p=2 is equivalent to the standard Euclidean metric.

3.3.3.8 XGBoost

The loss function to be optimized. ‘log_loss’ used in this model refers to binomial and multinomial deviance. Learning rate is 0.1 which shrinks the contribution of each tree by learning rate and number of boosting stages used is 100. The number of boosting stages to perform Gradient boosting is fairly robust to over-fitting so a large number usually results in better performance. There is a trade-off between learning rate and number of boosting stage. The function to measure the quality of a split are ‘friedman_mse’ for the mean squared error with improvement score by Friedman.

3.3.3.9 Artificial Neural Network (ANN)

The Sequential model allows us to build deep neural networks by stacking layers one on top of another. In neural networks literature, it’s common to talk about input layer, hidden layer and output layer. Input layers have 1000 nodes which we get by using tf-idf vector and activation function used is Rectified Linear Unit (ReLU). After that we put dropout layer sets to 0.3 which helps prevent overfitting and then we put two dense hidden layers with 500 nodes in each layer with activation function as ReLU and dropout layer sets as 0.3. The Dense function in Keras constructs a fully connected neural network layer, automatically initializing the weights as biases. The output of the this model is a single number because this is a binary classification.

So the output node as a vector with a single number (or simply a scalar) between 0 and 1 and activation function used is sigmoid.

We then compile the model with the compile function. Since we're building a binary 0/1 classifier, the loss function to minimize is binary_crossentropy and optimizer is 'adam', which use in order to minimize the loss function. Adam optimization is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. Now comes the final part of actually training the model using the fit function and we set number of times to go over the entire training data i.e., epochs to 3.

3.3.3.10 Convolutional Neural Network (CNN)

We use sequential model with embedding layer input dimension is 10000 which is size of the vocabulary and dimension of the dense embedding is 100. After that we add two convolution layer with number of filters is 512, size of filters is 8 and activation function is ReLU. After each convolution layer we add max pooling layer with size of the max pooling window is 4 and strides is 1 and also dropout layer with rate 0.2.

Two dense layer added with number of nodes is 100 in each layer and dropout rate is 0.2 with activation function is ReLU. Finally output layer with single node and activation function is sigmoid. For compile we use binary cross entropy with adam optimizer and number of epochs is 2.

3.3.3.11 Simple RNN

We use sequential model with embedding layer input dimension with 'maximum integer index size + 1' and dimension of the dense embedding is 300. We add one simple RNN layer with node size is 100, dropout rate is 0.2, recurrent dropout rate is 0.2 which is to drop for the linear transformation of the recurrent state and activation function is tanh. Output layer with single node and activation function is sigmoid. For compile we use binary cross entropy with adam optimizer and number of epochs is 3.

3.3.3.12 LSTM

We use sequential model with embedding layer with dimension of the dense embedding is 300. We add 3 lstm layer having 200 nodes in each layer, dropout rate is 0.2, recurrent dropout rate is

0.2 and activation function is tanh. Output layer with single node and activation function is sigmoid. For compile we use binary cross entropy with adam optimizer and number of epochs is 2.

3.3.3.13 GRU

We use sequential model with embedding layer with dimension of the dense embedding is 300. We add 5 GRU layer having 200 nodes in each layer, dropout rate is 0.2, recurrent dropout rate is 0.2 and activation function is tanh. Output layer with single node and activation function is sigmoid. For compile we use binary cross entropy with adam optimizer and number of epochs is 2.

3.3.3.14 Bidirectional LSTM

We use sequential model with embedding layer with dimension of the dense embedding is 300. We add 3 Bidirectional LSTM layer having 196 nodes in each layer, dropout rate is 0.2, recurrent dropout rate is 0.2 and activation function is tanh. Output layer with single node and activation function is sigmoid. For compile we use binary cross entropy with adam optimizer and number of epochs is 2.

3.3.3.15 Bidirectional Encoder Representations from Transformers (BERT)

We use TensorFlow Hub for BERT model and we use 'bert_multi_cased_L-12_H-768_A-12' model from keras layer. It uses L=12 hidden layers, a hidden size of H=768, and A=12 attention heads. BERT provides dense vector representations for natural language by using a deep, pre-trained neural network with the Transformer architecture. We convert text into lower case and then tokenize the tweets and convert tokens to ids after that we have added the input mask and the input type.

We have three layers of input and then we have used functional model and pooled output layer with dropout rate is 0.1. After that we add dense layer for predicting outputs with sigmoid activation function and compile with binary cross entropy for loss function and number of epochs used is 2.

3.3.3.16 Correlation Technique

After reading data we need to scale data for better result for that we use standard scaling method. Standard score of a sample x is calculated as: 'z = (x - u) / s' where u is the mean of the training and s is the standard deviation of the training samples. We use Pearson's correlation for all the features in dataset for finding the linear relationship and we use Spearman's correlation for finding non-linear relationship between them. Using PCA we find that some features contributing more in PC1 but not in PC2 so we can say that these feature are close in feature space.

3.4 Results and Analysis

Result for the model was calculated by Accuracy, Precision, Recall and F1 score measure. For better understanding we have given some basic information about accuracy, precision, recall and F1 score [48].

Accuracy is calculated by dividing the number of correct predictions (the corresponding diagonal in the matrix) by the total number of samples.

Precision is defined as what proportion of positive identifications was actually correct. In other words, how precise your model is out of those predicted positive, how many of them are actual positive.

$$\text{Precision} = \frac{\text{True Postive}}{\text{True Postive+False Positive}}$$

Recall is what proportion of actual positives was identified correctly. In other words how many of the actual Positives our model capture through labelling it as Positive (True Positive).

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

F1 Score is needed when you want to seek a balance between Precision and Recall.

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

We have calculated precision, recall and F1 score for class 0 and class 1 using weightage average. Also, we analyse result of all the models using stratified k-fold cross validation where k is ranging

from 2 to 10 which means if k = 4 then, 75% of the data is used for training data and 25% of the data is used for testing data. After that we find best split for each model to implementation.

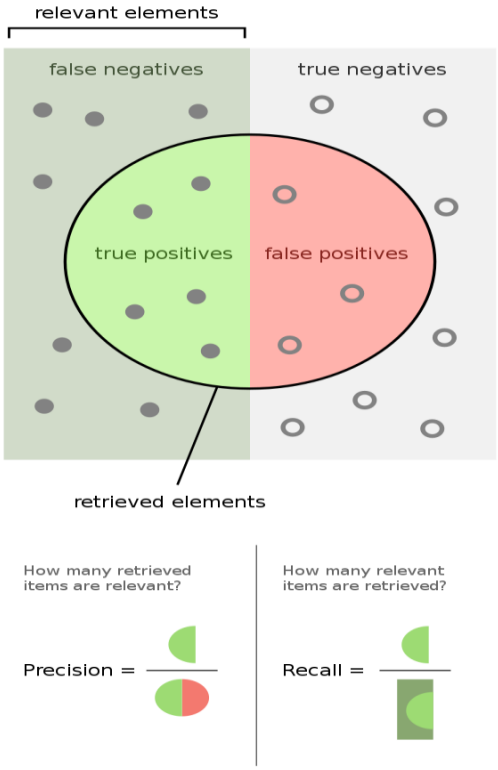

Figure 3.7: Precision and Recall

3.4.1 Model Evaluation

Table 3.1 below shows the accuracy, precision, recall and F1-score of all the different models which we have applied on the labelled tweets. Out of all the below models BERT performs better in all the evaluation metrics. So we use same BERT model with same hyperparameter on unlabelled tweets and evaluate depression score for all the tweets and then we find mean of all the tweets for a particular city and find the depression score of that city.

All Models	Accuracy	Precision	Recall	F1 score
Logistic Regression	93.22	0.93	0.93	0.93
Decision Tree	88.15	0.88	0.88	0.88
Random Forest	92.53	0.93	0.92	0.92
Multinomial Naive Bayes	86.15	0.87	0.86	0.86
Linear SVM	93.04	0.93	0.93	0.93
SVM (RBF)	93.97	0.94	0.94	0.94
KNN	73.46	0.79	0.73	0.76
XG Boost	88.41	0.88	0.87	0.87
ANN	93.34	0.93	0.93	0.93
CNN	92.88	0.93	0.92	0.93
Simple RNN	90.96	0.92	0.92	0.92
LSTM	93.77	0.94	0.94	0.94
GRU	93.93	0.94	0.94	0.94
Bidirectional LSTM	93.83	0.94	0.94	0.94
BERT	95.79	0.96	0.96	0.96

Table 3.1: Results of all the models

3.4.2 Correlation Evaluation

After applying BERT model on twitter data on 25 different cities in India we get the depression score of each cities. Figure 3.8 below shown give the visualisation relative depression score on map of India. Patna has highest depression whereas Thiruvananthapuram has lowest depression.

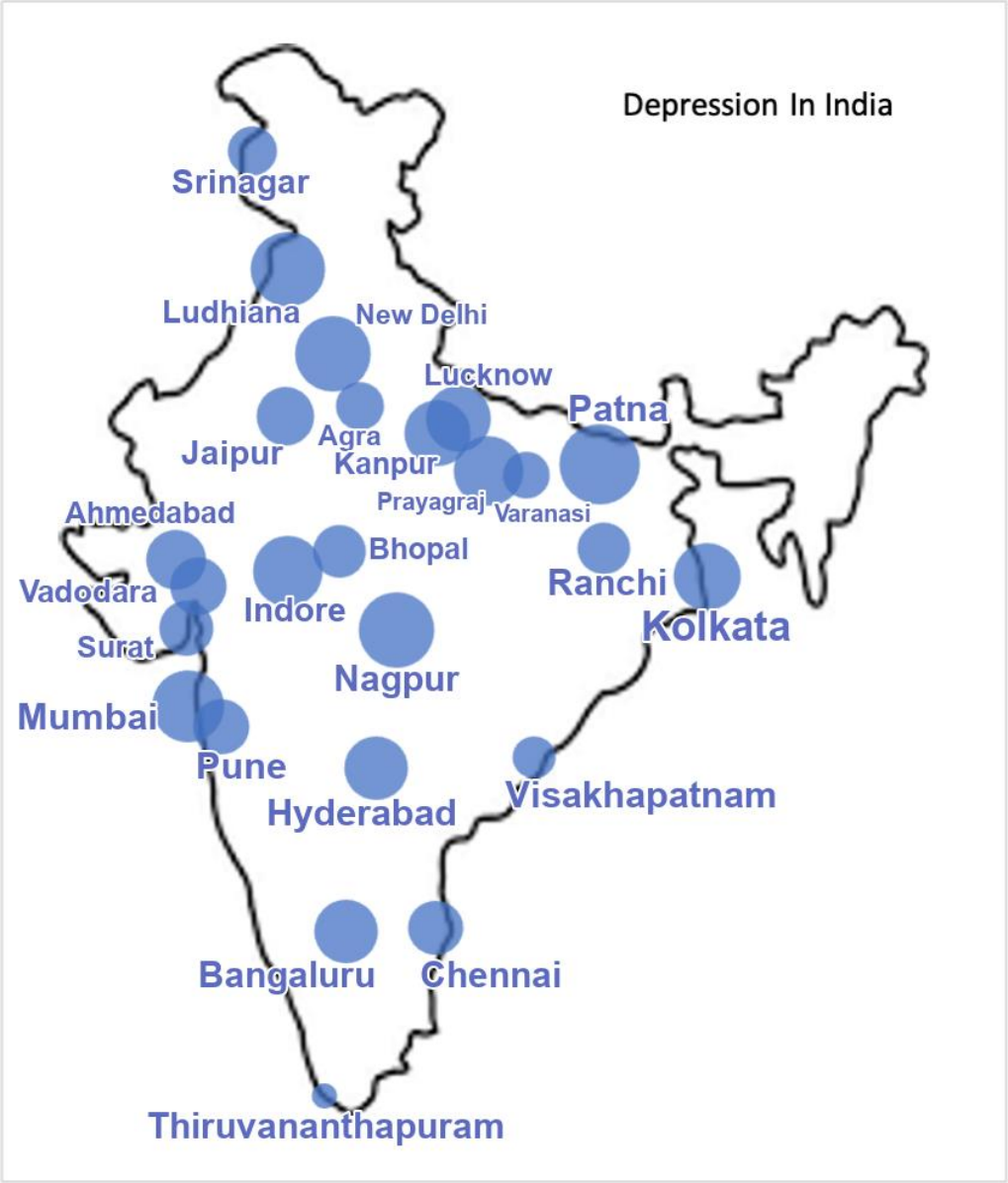

Figure 3.8: Depression in India

We have applied many different correlation technique for finding reasons for depression. We have shown correlation results using bar graph in figure 3.9. After analysing all the correlations we find out that number of never married person, number of person in a house, number of married female who is illiterate and number of non-workers have positive relation with depression and number of currently married, literacy rate, number of separated marriage, number of disabled person, number of female headed household have negative relation with depression.

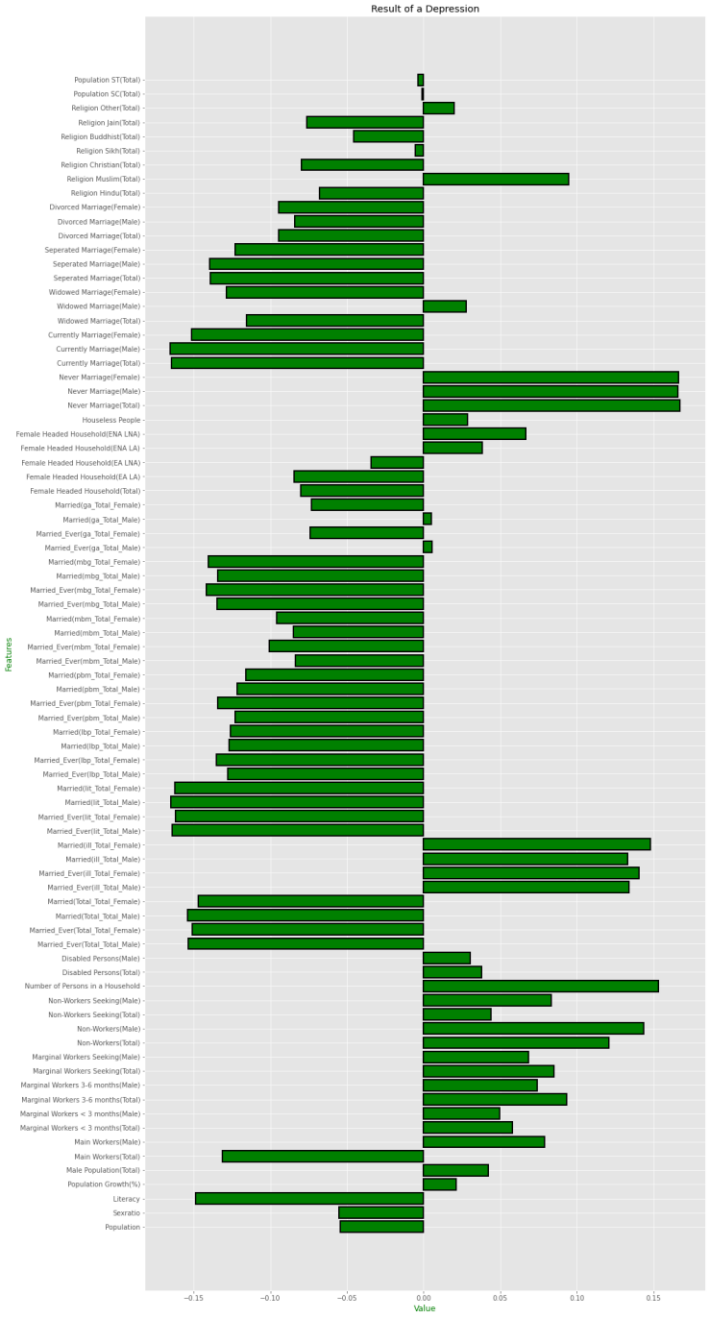

Figure 3.9: Correlation of different features with depression

Chapter 4

Conclusion and Future Work

This thesis presents a detailed study of demographic study of depression analysis in India using Social Media data. We analysis tweets and explored many different models for depression score but BERT outperforms from other models with accuracy of 95.79 and F1-score 0.96. Among all the 25 cities in India lowest depression score found in Thiruvananthapuram city and highest depression score found in Patna. We found out many different reasons for depression after analysing different correlations like number of unmarried person, number of person in house, number of non-workers are directly proportion to the depression and literacy rate, separated marriage, disabled person, female headed household are inversely proportional to the depression.

In future, we can convert this binary classification problem into multi class problem. In this research we have considered depression score between 0 to 1, further we can expand the range and analyse depression in more details. In this work we have analyzed only text data for depression. The work can be extended to depression analysis from multimodal social media contents (image, video, speech, text).

References

[1] https://www.who.int/health-topics/mental-health#tab=tab_2

[2] Coppersmith, G., Dredze, M. and Harman, C., 2014, June. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51-60).

[3] Li, I., Li, Y., Li, T., Alvarez-Napagao, S., Garcia-Gasulla, D. and Suzumura, T., 2020, December. What are we depressed about when we talk about covid-19: Mental health analysis on tweets using natural language processing. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 358-370). Springer, Cham.

[4] Schwartz, H.A., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., Kosinski, M. and Ungar, L., 2014, June. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 118-125).

[5] De Choudhury, M., Counts, S. and Horvitz, E., 2013, May. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference* (pp. 47-56).

[6] Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H. and Eichstaedt, J.C., 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*, pp.43-49.

[7] Shensa, A., Sidani, J.E., Dew, M.A., Escobar-Viera, C.G. and Primack, B.A., 2018. Social media use and depression and anxiety symptoms: A cluster analysis. *American journal of health behavior*, *42*(2), pp.116-128.

[8] Tadesse, M.M., Lin, H., Xu, B. and Yang, L., 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, *7*, pp.44883-44893.

[9] Matheson, F.I., Moineddin, R., Dunn, J.R., Creatore, M.I., Gozdyra, P. and Glazier, R.H., 2006. Urban neighborhoods, chronic stress, gender and depression. *Social science & medicine*, *63*(10), pp.2604-2616.

[10] Grover, S., Raju V, V., Sharma, A. and Shah, R., 2019. Depression in children and adolescents: A review of Indian studies. *Indian journal of psychological medicine*, *41*(3), pp.216-227.

[11] Pilania, M., Yadav, V., Bairwa, M., Behera, P., Gupta, S.D., Khurana, H., Mohan, V., Baniya, G. and Poongothai, S.J.B.P.H., 2019. Prevalence of depression among the elderly (60 years and above) population in India, 1997–2016: a systematic review and meta-analysis. *BMC public health*, *19*(1), pp.1-18.

[12] Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P. and Ansseau, M., 2003. Socioeconomic inequalities in depression: a meta-analysis. *American journal of epidemiology*, *157*(2), pp.98-112.

[13] Lynch, J.W., Kaplan, G.A. and Shema, S.J., 1997. Cumulative impact of sustained economic hardship on physical, cognitive, psychological, and social functioning. *New England Journal of Medicine*, *337*(26), pp.1889-1895.

[14] Weich, S. and Lewis, G., 1998. Poverty, unemployment, and common mental disorders: population based cohort study. *Bmj*, *317*(7151), pp.115-119.

[15] Lorant, V., Croux, C., Weich, S., Deliège, D., Mackenbach, J. and Ansseau, M., 2007. Depression and socio-economic risk factors: 7-year longitudinal population study. *The British journal of psychiatry*, *190*(4), pp.293-298.

[16] Lépine, J.P. and Briley, M., 2011. The increasing burden of depression. *Neuropsychiatric disease and treatment*, *7*(Suppl 1), p.3.

[17] Wittchen, H.U. and Jacobi, F., 2005. Size and burden of mental disorders in Europe—a critical review and appraisal of 27 studies. *European neuropsychopharmacology*, *15*(4), pp.357-376.

[18] Kessler, R.C. and Bromet, E.J., 2013. The epidemiology of depression across cultures. *Annual review of public health*, *34*, p.119.

[19] Freeman, A., Tyrovolas, S., Koyanagi, A., Chatterji, S., Leonardi, M., Ayuso-Mateos, J.L., Tobiasz-Adamczyk, B., Koskinen, S., Rummel-Kluge, C. and Haro, J.M., 2016. The role of socio-economic status in depression: results from the COURAGE (aging survey in Europe). *BMC public health*, *16*(1), pp.1-8.

[20] Haand, R. and Shuwang, Z., 2020. The relationship between social media addiction and depression: A quantitative study among university students in Khost, Afghanistan. *International Journal of Adolescence and Youth*, *25*(1), pp.780-786.

[21] Keles, B., McCrae, N. and Grealish, A., 2020. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth*, *25*(1), pp.79-93.

[22] Kessler, R.C., Amminger, G.P., Aguilar-Gaxiola, S., Alonso, J., Lee, S. and Ustun, T.B., 2007. Age of onset of mental disorders: a review of recent literature. *Current opinion in psychiatry*, *20*(4), p.359.

[23] Kim-Cohen, J., Caspi, A., Moffitt, T.E., Harrington, H., Milne, B.J. and Poulton, R., 2003. Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort. *Archives of general psychiatry*, *60*(7), pp.709-717.

[24] Lenhart, A., 2015. Teens, social media & technology overview 2015.

[25] Aydin, S., Koçak, O., Shaw, T.A., Buber, B., Akpinar, E.Z. and Younis, M.Z., 2021, April. Investigation of the effect of social media addiction on adults with depression. In *Healthcare* (Vol. 9, No. 4, p. 450). MDPI.

[26] Kirik, A., Arslan, A., Çetinkaya, A. and Mehmet, G.Ü.L., 2015. A quantitative research on the level of social media addiction among young people in Turkey. *International Journal of Sport Culture and Science*, *3*(3), pp.108-122.

[27] Karaci, A. and Piri, Z., 2017. Investigation of Facebook addiction of university students in terms of different variables: The case of Kastamonu University. *Kast. Educ. J*, *25*, pp.1547-1558.

[28] Çam, E. and Isbulan, O., 2012. A new addiction for teacher candidates: Social networks. *Turkish Online Journal of Educational Technology-TOJET*, *11*(3), pp.14-19.

[29] Altayef, H., 2018. *Investigation of social media addiction and usage purposes in terms of different variables* (Doctoral dissertation, Kastamonu University Institute of Science).

[30] Mayda, A., Yilmaz, M.U.A.M.M.E.R., Bolu, F., Dagli, S., Gercek, G., Teker, N., Tiryaki, S., Toygar, G., Turkarslan, M., Uslu, A. and Usturali, E., 2015. Internet addiction and Beck Depression Inventory in the university students at a student hostel. *KONURALP TIP DERGISI*, *7*(1).

[31] Kircaburun, K., 2016. Self-Esteem, Daily Internet Use and Social Media Addiction as Predictors of Depression among Turkish Adolescents. *Journal of Education and Practice*, *7*(24), pp.64-72.

[32] Aydm, B. and San, S.V., 2011. Internet addiction among adolescents: the role of self-esteem. *Procedia-Social and Behavioral Sciences*, *15*, pp.3500-3505.

[33] Bahrainian, S.A., Alizadeh, K.H., Raeisoon, M.R., Gorji, O.H. and Khazaee, A., 2014. Relationship of Internet addiction with self-esteem and depression in university students. *Journal of preventive medicine and hygiene*, *55*(3), p.86.

[34] Kim, J.E., Kang, H., Han, K. and Hong, Y.S., 2016. Internet use of korean adolescents: a test of causal model. *International Journal of Applied Engineering Research*, *11*(2), pp.1036-1141.

[35] Mei, S., Yau, Y.H., Chai, J., Guo, J. and Potenza, M.N., 2016. Problematic Internet use, well-being, self-esteem and self-control: Data from a high-school survey in China. *Addictive behaviors*, *61*, pp.74-79.

[36] Zhang, R., 2015. Internet dependence in Chinese high school students: Relationship with sex, self-esteem, and social support. *Psychological reports*, *117*(1), pp.8-25.

[37] Ayas, T. and Horzum, M., 2013. Relation between depression, loneliness, self-esteem and internet addiction. *Education*, *133*(3), pp.283-290.

[38] Reisoglu, I., Gedik, N. and GÖKTAŞ, Y., 2013. Relationship between pre-service teachers' levels of self-esteem, emotional intelligence and problematic internet use. *Egitim ve bilim-education and science*, *38*(170).

[39] Hormes, J.M., 2016. Under the influence of Facebook? Excess use of social networking sites and drinking motives, consequences, and attitudes in college students. *Journal of Behavioral Addictions*, *5*(1), pp.122-129.

[40] Griffiths, M.D., 2012. Facebook addiction: concerns, criticism, and recommendations—a response to Andreassen and colleagues. *Psychological reports*, *110*(2), pp.518-520.

[41] Cerniglia, L., Zoratto, F., Cimino, S., Laviola, G., Ammaniti, M. and Adriani, W., 2017. Internet Addiction in adolescence: Neurobiological, psychosocial and clinical issues. *Neuroscience & Biobehavioral Reviews*, *76*, pp.174-184.

[42] Griffiths, M., 1996. Behavioural addiction: an issue for everybody?. *Employee Councelling Today*, *8*(3), pp.19-25.

[43] Griffiths, M., 1998. Internet addiction: does it really exist?.

[44] Griffiths, M., 2005. A 'components' model of addiction within a biopsychosocial framework. *Journal of Substance use*, *10*(4), pp.191-197.

[45] Sepp Hochreiter and Jurgen Schmidhuber. "Long short-term memory". In: Neural computation 9.8 (1997), pp. 1735–1780.

[46] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[48] Richard Jizba. "Measuring search effectiveness". In: Creighton University Health Sciences Library and Learning Resources Center (2000).