

CLASSIFICATION OF SCHIZOPHRENIC DISEASE USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for
the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

HARI KANNAN (38110190)
HARIHARAN .R (38110191)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119

APRIL - 2022



SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **N HARI KANNAN (38110190)** and **HARIHARAN .R (38110191)** who carried out the project entitled “**CLASSIFICATION OF SCIZOPHRENIC DISEASE ANALYSIS USING MACHINE LEARNING**” under our supervision from November 2021 to April 2022.

Internal Guide
Mr. K. AMANDEEP M.E., Ph.D.,

Head of the Department
Dr. L. LAKSHMANAN M.E.,Ph.D.,
Dr.S.VIGNESHWARI, M.E.,Ph.D

Submitted for Viva-Voce Examination held on:

Internal Examiner

External Examiner

DECLARATION

We **N HARI KANNAN (38110190)** and **HARIHARAN .R (38110191)** hereby declare that the Project Report entitled “**CLASSIFICATION OF SCHIZOPHRENIC DISEASE ANALYSIS USING MACHINE LEARNING**” done by me under the guidance of **Ms.D.USHA NANDINI** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge our sincere thanks to **BOARD OF MANAGEMENT** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. SASIKALA M.E., Ph.D.**, Dean, School Of Computing, **Dr. S. VIGNESHWARI M.E., Ph.D.**, and **Dr. L. LAKSHMANAN M.E., Ph.D.**, Heads Of the Department, Department of Computer Science and Engineering for providing us the necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms.D.USHA NANDINI.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff of the **DEPARTMENT of COMPUTER SCIENCE & ENGINEERING** who were helpful in many ways for the completion of the project.

ABSTRACT

Schizophrenia is a mental issue that typically shows up in late immaturity or early adulthood. Described by dreams, mental trips, and other subjective troubles, schizophrenia can regularly be a long lasting battle. Schizophrenia is a mental issue wherein practical and basic cerebrum systems are upset. Old style organize investigation has been utilized by numerous scientists to measure mind systems and to examine the system changes in schizophrenia, however lamentably measurements utilized in this old style technique exceptionally rely upon the systems' thickness and weight; the examinations made by this strategy are one-sided. Customary techniques either need enough memorable information or to keep the customary checking on persistent exercises for recognizable proof of a patient related with a mental ailment infection. Strategy: In request to address this issue, we propose a philosophy to characterize the patients related with constant mental sickness maladies (for example Uneasiness, Depression, Bipolar, and Consideration Deficit Hyperactivity Disorder) in view of the information removed from the Reddit, an outstanding arrange network stage. The proposed technique is utilized through Co-training technique by joining the discriminative intensity of generally utilized classifiers.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	v
	LIST OF FIGURES	viii
	LIST OF ABBREVIATIONS	ix
1	INTRODUCTION	1
	1.1 OVERVIEW	3
2	LITERATURE SURVEY	4
3	METHODOLOGY	
	3.1 AIM OF THE PROJECT	7
	3.2 HARDWARE REQUIREMENTS	8
	3.3 SOFTWARE REQUIREMENTS	8
	3.4 OVERVIEW OF PYTHON	9
	3.5 SYSTEM DESIGN	10
	3.5.1 System Architecture	10
	3.5.2 Module overview	11
	3.5.2.1 Loading Live Data	12
	3.5.2.2 Data Preprocessing	13
	3.5.2.3 Building the Model	14
	3.5.2.4 Training the Model	15
	3.6 RANDOM FOREST ALGORITHM	16
	3.7 DECISION TREE ALGORITHM CLASSIFICATION	19
	3.8 RAD	25
	3.9 BENEFITS	28

4.	RESULT AND DISCUSSION	
	4.1 DECISION TREE BOUNDARIES	30
	4.2 LOGISTIC REGRESSION	30
	4.3 DATA PREPROCESSING	31
	4.4 TEXT TOKENIZATION	32
	4.5 CLASSIFICATION	33
5.	SUMMARY AND CONCLUSION	
	5.1 SUMMARY	34
	5.2 CONCLUSION	35
	5.3 FUTURE WORK	35
	REFERENCES	37
	APPENDIX	
	A.SOURCE CODE	39
	B.PUBLICATION	46

LIST OF FIGURES

FIGURE NO.	NAME OF THE FIGURE	PAGE NO.
3.1	ARCHITECTURE DIAGRAM	11
3.2	RANDOM FOREST	17
3.3	DECISION TREE	18
3.4	ARCHITECTURE OF DECISION TREE	20
3.5	TREE DIAGRAM	22
3.6	RAD MODEL	25
4.1	PREPROCESS DATA	31
4.2	KEYWORD VECTORS	32
4.3	PREDCTION	33
4.4	MODEL ACCURACY	33

LIST OF ABBREVIATIONS

RAD	Rapid Application Development
SDLC	Software Development Life Cycle
NLP	Natural Language programming
NE	Named Entity
RNN	Recurrent Neural Network

CHAPTER 1

INTRODUCTION

Schizophrenia is a dangerous neurological sickness, which seems in roughly 1% of the adult population around the globe. It affects remembrance, awareness, linguistic and executive functions of the brain. According to World Health Organization (WHO) stated that the worldwide incidence of global Schizophrenia is just about 20 million people. Among that 69% percentage of Schizophrenia disorder patients are not obtaining proper meditations. There are various causes for Schizophrenia disorders.

Any research does not distinguish one particular aspect. Hereditary and environmental influences perform a major role in its diagnosis. It is normally developed in the adulthood under the age of 26. The patients which are affected by this disease will have intellectual deficiencies, including reading disorders. They will have difficulty to select suitable phrases while they are talking and composing. It modifies patients' observation, points of view, and conduct as confirm by fantasies, dreams, muddled discourse or conduct, social withdrawal, and differed intellectual deficiencies. Fig 1.

Schizophrenia Word Cloud Schizophrenia Discourse (SD) is a mental ailment, which is the highly predominant neurodegenerative disease in this globe. It is a radical and dense mental illness influencing susceptible peoples, whose pervasiveness is projected to twice by 2030. Diagnosis and medication of neurological Discourse are emphases in brain study. Despite the fact that meds have been created to treat schizophrenia, they don't work for all patients, and there is proof that a few types of schizophrenia seem, by all accounts, to be treatment safe. This just underlines the significance of leading fruitful schizophrenia clinical preliminaries.

According to the artificial intelligence study, social media text analysis has been a leading area of study in the past two to three decades that has been commonly applied in wide array of research such as healthcare, smart home, surveillance systems, human-computer interaction, gaming and so on. With an exceptional development of web-based social networking, huge number of individuals deliberately share a lot of information by communicating their states of mind, sentiments, feelings,

and day by day battles with emotional well-being issues via web-based networking media stages like Twitter.

This offers open doors for new comprehension of these networks. With the improvement of Internet utilization, individuals have begun to impart their encounters and difficulties to emotional wellness issue through online discussions, small scale web journals or tweets. Their online exercises roused numerous specialists to present new types of potential human services arrangements and techniques for early melancholy recognition frameworks. Utilizing distinctive Natural Language Processing (NLP) methods and content characterization draws near, they attempted to prevail in a better improvement.

The singular aim of this application is to develop a machine learning especially deep learning approach, named Long-Short Term Memory (LSTM) plus Convolutional Neural Network (CNN), for detecting Schizophrenia disorders using features obtained from a different dataset from kaggle.com, namely the Schizophrenia Discourse dataset. The proposed LSTM+CNN is competent to handle high-ranking dimensional data by expanding the model interpretability and discovering the highly appropriate features while dispensing with the typical inequitable class division. Natural language processing (NLP) has gained significantly from the revival of neural networks (NNs), expected to their superior execution with fewer necessary of contrived features.

Convolutional neural network (CNN) and recurrent neural network (RNN) are the two major types Deep neural network architecture. Gating systems have been established to improve certain constraints of the basic recurrent neural network, causing in two fundamental recurrent neural network types: long short-term memory (LSTM) and gated recurrent unit (GRU). This application is therefore dedicated on approaches to categorize symptoms of Schizophrenia Discourse in posted tweets as early as feasible established on deep learning methodology.

This paper content is organized into six divisions. The first division contains the Introduction section. The second division specifies about social network. The third division illuminates related work. Our proposed system is deliberated in the fourth division which covers the datasets, data analysis procedure, the proposed LSTM+CNN approach. The experiment results are presented in the fifth division and

conclusions are illustrated in the final segment. The final segment also contains the discussion as well as the future work in that direction.

1.1 OVERVIEW

The social media network has become immeasurable and reputation along with all the service industries today. Today most of the people around the world are using the social media network to share their individual, qualified interest and condition consciousness. Social media networks are the collection of collaborative web-based technology where interested communities produce and share expertise, concepts, intelligence, specialized interests and other information through virtual communities using online communication channels. There are various types of social medias like wikis, Social Networks, Media Sharing Networks, Discussion Forums, Bookmarking & Content Curation Networks, Social Review Websites, Community Blogging & Publishing Websites, Image & Video Hosting Networks, Sharing Economy Networks, Social Shopping Networks, Interest-Based Networks, Anonymous Social Networks and many more. The popular social networks are Facebook, Twitter, LinkedIn etc. The social networks are web based social media applications which are constantly evolving. The communities uses the social network to connect people and share knowledge through human-to-human interaction. It is also called as relationship networks where likeminded people can share ideas and knowledge. The popular image and video hosting websites are Flickr, photobucket, Instagram, Snapchat, YouTube. Vimeo etc. It is also called as Media Sharing Websites. These media sharing networks can be used to share photos, images, videos, live videos, audios and other medias which can be accessed from anywhere and anytime around the world. The image and video hosting websites can be used to promote brand recognition, lead creation etc. These networks mostly promote user generated content. The widespread Discussion

CHAPTER 2

LITERATURE SURVEY

Joseph Kambeitz, Stefan Leucht, Lana Marija Kambeitz-Illankovic, Stephen James Wood Detecting Neuroimaging Biomarkers for Schizophrenia In the year 2015. Schizophrenia is a mental issue that typically shows up in late immaturity or early adulthood. Described by dreams, mental trips, and other subjective troubles, schizophrenia can regularly be a long lasting battle. Schizophrenia is a mental issue wherein practical and basic cerebrum systems are upset. Old style organize investigation has been utilized by numerous scientists to measure mind systems and to examine the system changes in schizophrenia, however lamentably measurements utilized in this old style technique exceptionally rely upon the systems' thickness and weight; the examinations made by this strategy are one-sided.

Kimberly McManus, Emily K. Mallory, Rachel L. Goldfeder, MS, Winston A. Haynes BA Mining Twitter Data to Improve Detection of Schizophrenia In the year 2016. Customary techniques either need enough memorable information or to keep the customary checking on persistent exercises for recognizable proof of a patient related with a mental ailment infection. Strategy: In request to address this issue, They proposed a philosophy to characterize the patients related with constant mental sickness maladies (for example Uneasiness, Depression, Bipolar, and Consideration Deficit Hyperactivity Disorder) in view of the information removed from the Reddit, an outstanding arrange network stage. The proposed technique is utilized through Co-training technique by joining the discriminative intensity of generally utilized classifiers.

Jonathan D. Tatum, IStanford University, Stanford, CA Multivariate Pattern Recognition Studies In the year 2015 Individuals who suffer from schizophrenia comprise 1 percent of the United States Population and are four times more likely to die of suicide than the general US population. Identification of at-risk individuals with schizophrenia is challenging when they do not seek treatment. Microblogging platforms allow users to share their thoughts and emotions with the world in short snippets of text. In this work. They leveraged a large corpus of Twitter posts and machine-learning methodologies to detect individuals with schizophrenia. Using features from tweets.

Wolfram Hinzen and Joana Rossell The linguistics of schizophrenia thought disturbance as language pathology across positive symptoms. In the Year 2015 Emoticon use, posting time of day, and dictionary terms. They trained, built, and validated several machine learning models. The Support Vector machine model achieved the best performance with 92% precision and 71% recall on the held test set. Additionally, They built a web application that dynamically displays summary Statistics between cohorts. This enables outreach to undiagnosed individuals. Improved physician diagnoses, and destigmatization of schizophrenia. This Method hypothesize that linguistic organization in the schizophrenic brain plays a more central role in the pathogenesis of this disease than commonly supposed. Against the standard view, that schizophrenia is a disturbance of thought or selfhood, They argue that the origins of the relevant forms of thought and self hood at least partially depend on language. The view that they do not is premised by a theoretical conception of language that They identified as 'Cartesian' and contrast with a recent 'un Cartesian' model. This linguistic model empirically argues for both (i) a one-to-one correlation between human specific thought or meaning and forms of grammatical organization, and (ii) an integrative and codependent view of linguistic cognition and its sensory motor dimensions. Core dimensions of meaning mediated by grammar on this model specifically concern forms of referential and propositional meaning. A breakdown of these is virtually definitional of core symptoms. Within this model the three main positive symptoms of schizophrenia falling to place as failures in language-mediated forms of meaning, manifest either as a disorder of speech perception (Auditory Verbal Hallucinations), abnormal speech production running without feedback control (Formal Thought Disorder), or production of abnormal linguistic content (Delusions). The hypothesis make settable predictions for the language profile of schizophrenia across symptoms; it simplifies the cognitive neuro psychology

Deanna Greenstein , James D. Malley , Brian Weisinger , Liv Clasen and Nitin Gogtay Multivariate machine learning methods can be used to classify groups of schizophrenia patients and controls using structural magnetic resonance imaging (MRI) In the year 2014. However, machine learning methods to date have not been extended beyond classification and contemporaneously applied in a meaningful way to clinical measures. They hypothesized that brain measures would classify groups, and that increased likelihood of being classified as a patient using regional

brain measures would be positively related to illness severity, developmental delays, and genetic risk. Methods: Using 74 anatomic brain MRI sub regions and Random Forest(RF) ,a machine learning method, They classified 98 childhood on set schizophrenia(COS) patients and 99 age ,sex, and ethnicity-matched healthy controls. They also used RF to estimate the probability of being classified as a schizophrenia patient based on MRI measures. Then explored relationships between brain-based probability of multivariate machine learning methods can be used to classify groups of schizophrenia patients and controls using structural magnetic resonance imaging(MRI).

CHAPTER 3

METHODOLOGY

3.1 AIM OF THE PROJECT

To implement an end-to-end extensible framework for extracting the information about Schizophrenia **symptoms** delusions. The hallucinations, catatonic or disorganized .behavior, apathy, reduced thought fluidity, disperse and unproductive language Using Machine Learning.

One of the primary choices we face for our e-voting project implementations is -Which development methodology should we use?! There are various software development methodologies are available in the ICT industry. They are

- Agile Software Development
- Crystal Methods
- Dynamic Systems Development Model (DSDM)
- Extreme Programming (XP)
- Feature Driven Development (FDD)
- Joint Application Development (JAD)
- Lean Development (LD)
- Rapid Application Development (RAD)
- Rational Unified Process (RUP)
- Scrum
- Spiral
- Systems Development Life Cycle (SDLC)
- Waterfall (a.k.a. Traditional)

Among all these methodologies, we have decided to compare the two most popular and widely used methodologies:

- Waterfall Model
- RAD Model

3.2 HARDWARE REQUIREMENT

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It shows what the system does and not how it should be implemented.

PROCESSOR	:	PENTIUM IV 2.6 GHz, Intel Core 2 Duo.
RAM	:	512 MB DD RAM
MONITOR	:	15" COLOR
HARD DISK	:	40 GB

3.3 SOFTWARE REQUIREMENT

The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the team and tracking the team's progress throughout the development activity.

Technologies	:	Python, SQL sever
Operating System	:	Windows 10
IDE	:	Jupyter

3.4 OVERVIEW OF PYTHON

Python coding style comprises physical lines as well as logical lines or statements. A physical line in a Python program is a sequence of characters, and the end of the line terminates the line sequence as opposed to some other languages, such as C and C++ where a semi-colon is used to mark the end of the statement. A logical line, on the other hand, is composed of one or more physical lines. The use of a semi-colon is not prohibited in Python, although it's not mandatory. The NEWLINE token denotes the end of the logical line. A logical line that only contains spaces, comments, or tabs are called blank lines and they are ignored by the interpreter.

As we saw that in Python, a new line simply means that a new statement has started. Although, Python does provide a way to split a statement into a multiline statement or to join multiple statements into one logical line. This can be helpful to increase the readability of the statement. Following are the two ways to split a line into two or more lines:

Explicit Line Joining

In explicit line joining, we use a backward slash to split a statement into a multiline statement.

Implicit Line Joining

Statements that reside inside [], {}, or () parentheses can be broken down into two or more physical lines without using a back slash.

Multiple Statements on a Single Line

In Python, it is possible to club multiple statements in the same line using a semi-colon; however, most programmers do not consider this to be a good practice as it reduces the readability of the code.

3.5 SYSTEM DESIGN

Preprocessing and preparing the right kind of data for schizophrenic disease analysis. Text tokenizing is done on the data set. Determining the vector phrases. Speech tagging is done on the tokens. Removing duplicates through shallow parsing. Then classification is done based on the machine learning algorithm used. Rephrase all tweeter seeds as phrases or unigrams. Delete all retweet handlers. Delete all the URL from each tweets. Delete all Non-ASCII characters and words from the Schizophrenia discourse dataset and NonSchizophrenia discourse dataset. Break Usernames and Hashtags. Delete any punctuations from the hashtags. Delete all punctuations from the tweets. Load the prebuilt datasets and remove those stop words from datasets. Delete all leading and trailing spaces. Step 10: Delete all numbers from the tweets. Delete all multiple in between text white spaces from the tweets.

Input

The Data sets collected from the the Scizophrenic and non Schizophrenic patients.

Output

Based on the prediction and accuracy to classify the disease.

3.5.1 System Architecture

Text tokenizing is done on the data set. Determining the vector phrases. Speech tagging is done on the tokens. Removing duplicates through shallow parsing. Then classification is done based on the machine learning algorithm used. Rephrase all tweeter seeds as phrases or unigrams. Delete all retweet handlers. Delete all the URL from each tweets. Delete all Non-ASCII characters and words from the Schizophrenia discourse dataset and NonSchizophrenia discourse dataset. Break Usernames and Hashtags. Delete any punctuations from the hashtags. Delete all punctuations from the tweets. Load the prebuilt datasets and remove those stop

words from datasets .Delete all leading and trailing spaces. Step 10: Delete all numbers from the tweets. Delete all multiple in between text white spaces from the tweets.

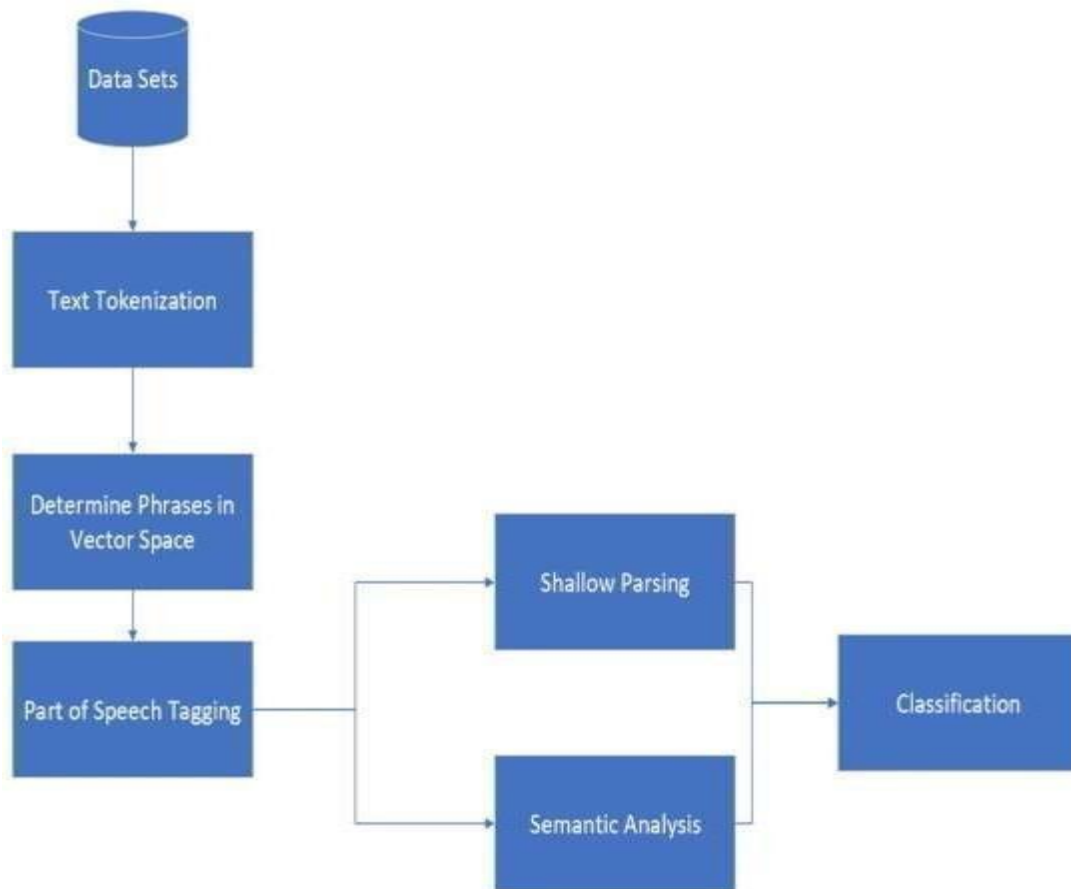


Fig.3.1: Architecture Diagram

3.5.2 Module Overview

- Loading live data
- Data Preprocessing
- Building the model
- Training Model

3.5.2.1 Loading Live Data

Text mining is the application of natural language processing techniques and analytical methods to text data in order to derive relevant information. Twitter data constitutes a rich source that can be used for capturing information about any topic imaginable. This data can be used in different use cases such as finding trends related to a specific keyword, measuring brand sentiment, and gathering feedback about new products and services.

Twitter is a gold mine of data. Unlike other social platforms, almost every user's tweets are completely public and pullable. This is a huge plus if you're trying to get a large amount of data to run analytics on. Twitter data is also pretty specific. Twitter's API allows you to do complex queries like pulling every tweet about a certain topic within the last twenty minutes or pull a certain user's non-retweeted tweets.

This data can be used in different use cases such as finding trends related to a specific keyword, measuring brand sentiment, and gathering feedback about new products and services.

In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret. Follow the steps below to get all 4 elements:

Create a twitter account if you do not already have one.

Go to <https://apps.twitter.com/> and log in with your twitter credentials.

Click "Create New App"

Fill out the form, agree to the terms, and click "Create your Twitter application"

In the next page, click on "API keys" tab, and copy your "API key" and "API secret".

Scroll down and click "Create my access token", and copy your "Access token" and "Access token secret".

3.5.2.2 *Data Preprocessing*

Data preprocessing is an important step to prepare the data to form a Schizophrenic Discourse model. There are many important steps in data preprocessing, such as data cleaning, data transformation, and feature selection. Schizophrenic Discourse Data cleaning and transformation are methods used to remove outliers and standardize the data so that they take a form that can be easily used to create a model. A Schizophrenic Discourse data set may contain hundreds of variables (descriptors); however, many of these variables will contain redundant data. In order to simplify the dimensionality of the model, it is important to select only variables that contain unique and important information. Data mining procedures can be used to remove variables that do not contribute to the Schizophrenic Discourse model.

We perform a series of preprocessing steps before feature extraction and classification is done. Named entity (NE) recognition, coreference resolution, and dependence parsing are performed using the library. Schizophrenic Discourse data are generally

Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

Noisy: containing errors or outliers

Inconsistent: containing discrepancies in codes or names

Tasks in data preprocessing

Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

Data integration: using multiple databases, data cubes, or files.

Data transformation: normalization and aggregation.

Data reduction: reducing the volume but producing the same or similar analytical results.

Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Many words in Schizophrenic Discourse dataset recur very frequently but are essentially meaningless as they are used to join words together in a sentence. It is commonly understood that stop words do not contribute to the context or content of textual documents. Due to their high frequency of occurrence, their presence in text mining presents an obstacle in understanding the content of the documents. Stop words are very frequently used common words like 'and', 'are', 'this' etc. They are not useful in classification of documents. So they must be removed. However, the development of such stop words list is difficult and inconsistent between textual sources. This process also reduces the text data and improves the Schizophrenic Discourse system performance. Every text Schizophrenic Discourse dataset deals with these words which are not necessary for text mining applications.

3.5.2.3 Building the Model

A LSTM network is a kind of recurrent neural network. A recurrent neural network is a neural network that attempts to model time or sequence dependent behaviour – such as language, stock prices, electricity demand and so on. This is performed by feeding back the output of a neural network layer at time t to the input of the same network layer at time $t + 1$.

An LSTM network is a recurrent neural network that has LSTM cell blocks in place of our standard neural network layers. These cells have various components called the input gate, the forget gate and the output gate.

This memory allows the network to learn long-term dependencies in a sequence which means it can take the entire context into account when making a prediction, whether that be the next word in a sentence, a sentiment classification, or the next temperature measurement. A RNN is designed to mimic the human way of processing sequences: we consider the entire sentence when forming a response instead of words by themselves.

3.5.2.4 *Training the Model*

The Long Short-Term Memory network, or LSTM network, is a recurrent neural network that is trained using Backpropagation Through Time and overcomes the vanishing gradient problem. As such, it can be used to create large recurrent networks that in turn can be used to address difficult sequence problems in machine learning and achieve state-of-the-art results. Instead of neurons, LSTM networks have memory blocks that are connected through layers.

A block has components that make it smarter than a classical neuron and a memory for recent sequences. A block contains gates that manage the block's state and output. A block operates upon an input sequence and each gate within a block uses the sigmoid activation units to control whether they are triggered or not, making the change of state and addition of information flowing through the block conditional.

There are three types of gates within a unit:

Forget Gate: conditionally decides what information to throw away from the block.

Input Gate: conditionally decides which values from the input to update the memory state.

Output Gate: conditionally decides what to output based on input and the memory of the block.

Each unit is like a mini-state machine where the gates of the units have weights that are learned during the training procedure.

LSTMs can be used to model univariate time series forecasting problems. These are problems comprised of a single series of observations and a model is required to learn from the series of past observations to predict the next value in the sequence. The LSTM model will learn a function that maps a sequence of past observations as input to an output observation. As such, the sequence of observations must be transformed into multiple examples from which the LSTM can learn.

The Long Short Term Memory neural network is a type of a Recurrent Neural Network (RNN). RNNs use previous time events to inform the later ones. For example, to classify what kind of event is happening in a movie, the model needs to use

information about previous events. RNNs work well if the problem requires only recent information to perform the present task. If the problem requires long term dependencies.

As such, it can be used to create large recurrent networks that in turn can be used to address difficult sequence problems in machine learning and achieve state-of-the-art results. Instead of neurons, LSTM networks have memory blocks that are connected through layers. A block has components that make it smarter than a classical neuron and a memory for recent sequences

3.6 RANDOM FOREST ALGORITHM

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm\ Random forest is a . The "forest" it builds, is an ensemble of decision trees, usually trained with the -bagging| method. The general idea of the bagging method is that a combination of learning models increases the overall result.

We can understand the working of Random Forest algorithm with the help of following steps –

- **Step 1** – First, start with the selection of random samples from a given dataset.
- **Step 2** – Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- **Step 3** – In this step, voting will be performed for every predicted result.
- **Step 4** – At last, select the most voted prediction result as the final prediction - result.

The following diagram will illustrate its working –

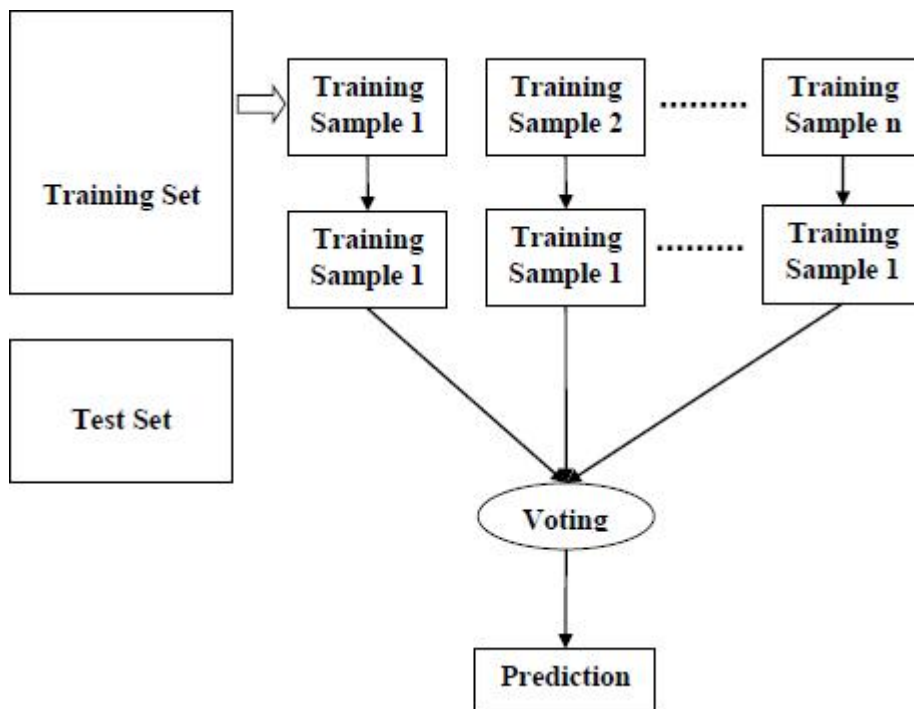


Fig.3.2: Random Forest

RANDOM FOREST ALGORITHM WORKING

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the -baggingl method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Put simply random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:

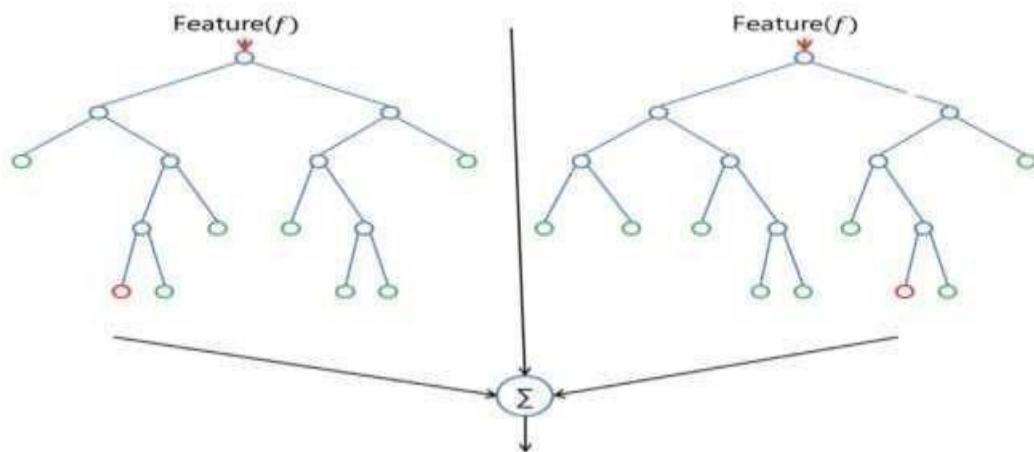


Fig.3.3: Decision Tree

Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor.

Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree)

3.7 DECISION TREE ALGORITHM CLASSIFICATION

Decision Tree is a Supervised learning technique that can be used for both for Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:

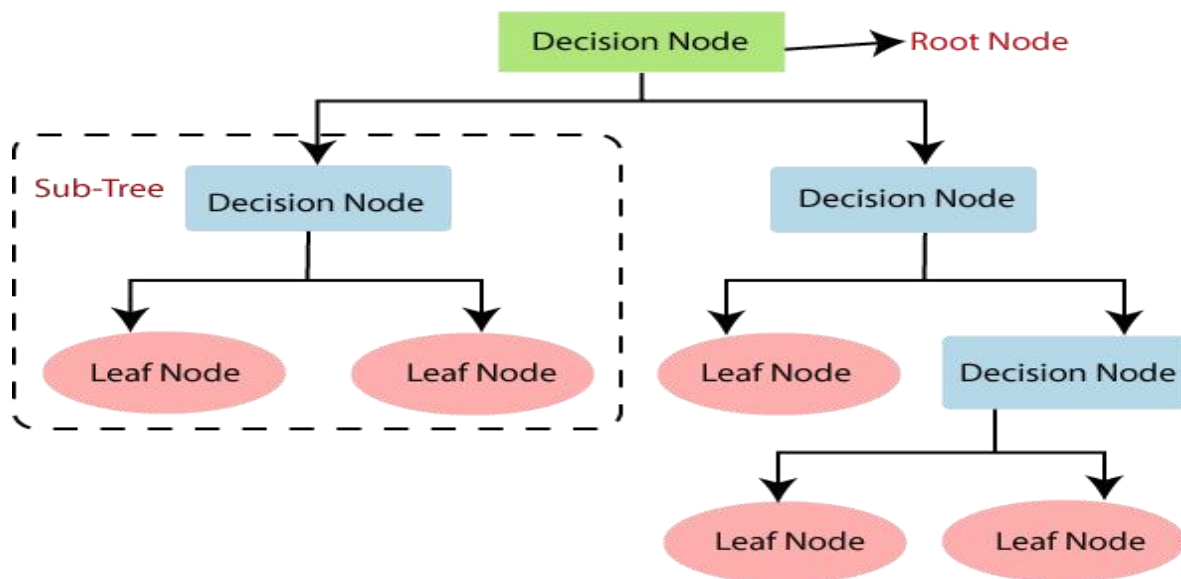


Fig.3.4: Architecture of Decision Tree

WHY USE DECISION TREES

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

DECISION TREE TERMINOLOGIES

Root Node

Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

Leaf Node

Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

Splitting

Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

Branch/Sub Tree

A tree formed by splitting the tree.

Pruning

Pruning is the process of removing the unwanted branches from the tree.

Parent/Child node

The root node of the tree is called the parent node, and other nodes are called the child nodes.

DECISION ALGORITHM WORKING

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the decision tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of the dataset created.

Example: Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:

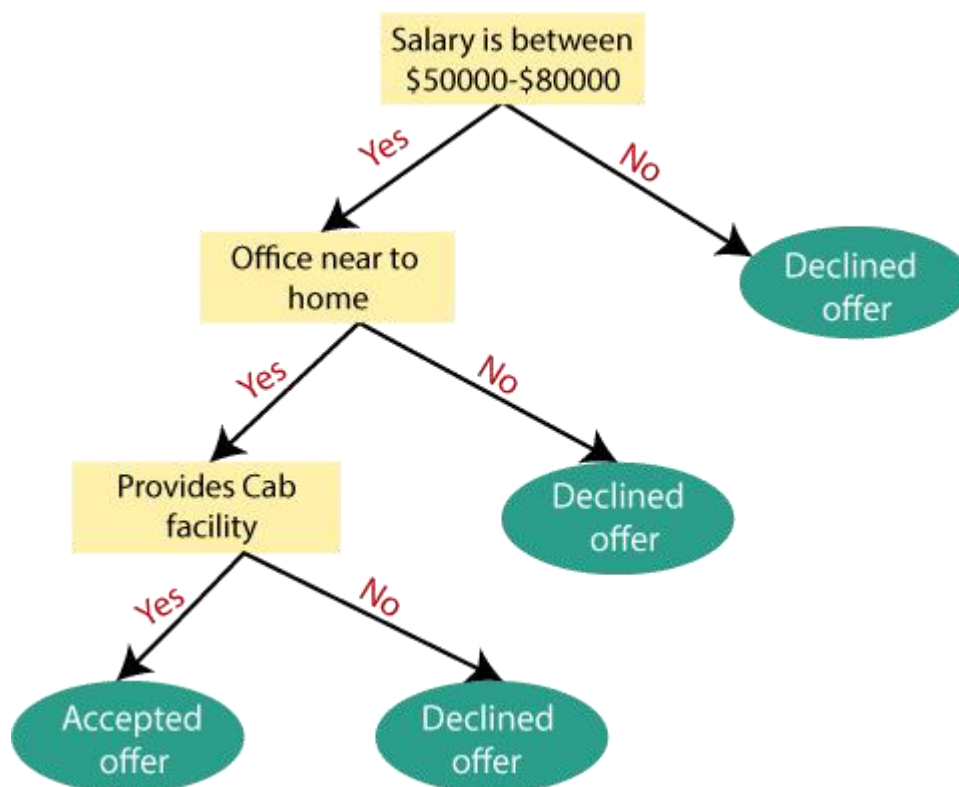


Fig.3.5: DecisionTree Diagram

Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM. By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- Information Gain

- Gini Index

Information Gain

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)]

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= $-P(\text{yes})\log_2 P(\text{yes}) - P(\text{no}) \log_2 P(\text{no})$

Where,

- S= Total number of samples
- P(yes)= probability of yes
- P(no)= probability of no

Gini Index:

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1 - \sum P_j^2$$

Pruning: Getting an Optimal Decision tree

Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of tree pruning technology used:

- Cost Complexity Pruning
- Reduced Error Pruning.

Advantages of the Decision Tree

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

Disadvantages of the Decision Tree

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue, which can be resolved using the Random Forest algorithm.
- For more class labels, the computational complexity of the decision tree may increase.

Python Implementation of Decision Tree

Now we will implement the Decision tree using Python. For this, we will use the dataset "user_data.csv," which we have used in previous classification models. By

using the same dataset, we can compare the Decision tree classifier with other classification models such as KNN SVM, LogisticRegression, etc.

Steps will also remain the same, which are given below:

- Data Pre-processing step
- Fitting a Decision-Tree algorithm to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.

3.8 RAD

Agile is an iterative, team-based approach to development. This method highlights the speedy delivery of an software application in comprehensive functional components.

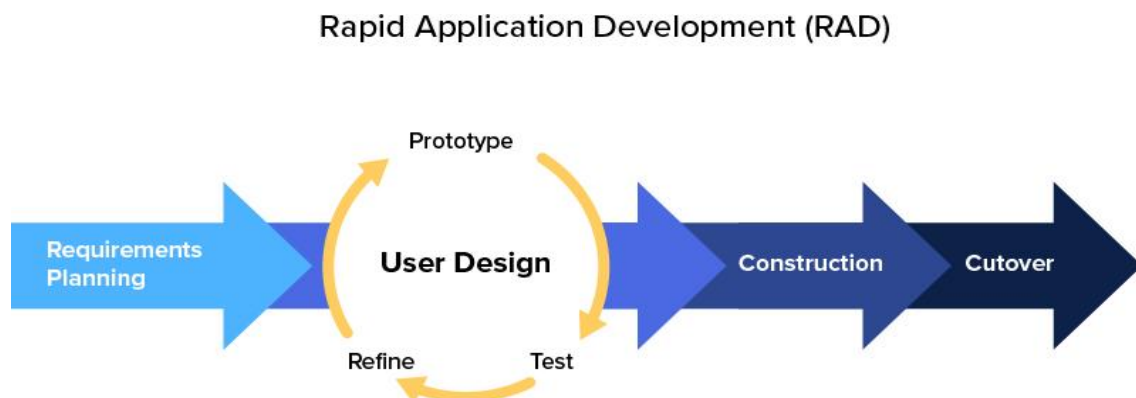


Fig.3.6: RAD Model

Requirements planning

This phase is equivalent to a project scoping meeting. Although the planning phase is condensed compared to other project management methodologies, this is a critical step for the ultimate success of the project.

During this stage, web application programmers, clients (software users), and team members communicate to determine the goals and expectations for the project as well as current and potential issues that would need to be addressed during the build.

A basic breakdown of this stage involves:

Researching the current problem

Defining the requirements for the project

Finalizing the requirements with each stakeholder's approval

It is important that everyone has the opportunity to evaluate the goals and expectations for the project and weigh in. By getting approval from each key stakeholder and web application programmer, teams can avoid miscommunications and costly change orders down the road.

User design

Once the project is scoped out, it's time to jump right into development, building out the user design through various prototype iterations.

This is the meat and potatoes of the RAD methodology—and what sets it apart from other project management strategies. During this phase, clients work hand in hand with web application programmers to ensure their needs are being met at every step in the design process. It's almost like customizable software development where the users can test each prototype of the product, at each stage, to ensure it meets their expectations.

All the bugs and kinks are worked out in an iterative process. The web application programmer designs a prototype, the client (user) tests it, and then they come together to communicate on what worked and what didn't.

This method gives web application programmers the opportunity to tweak the model as they go until they reach a satisfactory design.

Both the software web application programmers and the clients learn from the experience to make sure there is no potential for something to slip through the cracks.

Rapid construction

Phase 3 takes the prototypes and beta systems from the design phase and converts them into the working model.

Because the majority of the problems and changes were addressed during the thorough iterative design phase, web application programmers can construct the final working model more quickly than they could by following a traditional project management approach.

The phase breaks down into several smaller steps:

Preparation for rapid construction

Program and application development

Coding

Unit, integration, and system testing

The software development team of programmers, coders, testers, and web application programmers work together during this stage to make sure everything is working smoothly and that the end result satisfies the client's expectations and objectives.

This third phase is important because the client still gets to give input throughout the process. They can suggest alterations, changes, or even new ideas that can solve problems as they arise.

Cutover

This is the implementation phase where the finished product goes to launch. It includes data conversion, testing, and changeover to the new system, as well as user

training. All final changes are made while the coders and clients continue to look for bugs in the system.

3.9 Benefits

RAD is one of the most successful software development programs available today, with numerous benefits for both software development teams as well as their clients.

Here are just a few advantages:

RAD lets you break the project down into smaller, more manageable tasks.

The task-oriented structure allows project managers to optimize their team's efficiency by assigning tasks according to members' specialties and experience.

Clients get a working product delivered in a shorter time frame.

Regular communication and constant feedback between team members and stakeholders increases the efficiency of the design and build process.

PROS

The following are the advantages of Random Forest algorithm –

- It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
- Random forests work well for a large range of data items than a single decision tree does.
- Random forest has less variance than single decision tree.
- Random forests are very flexible and possess very high accuracy.
- Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling.
- Random Forest algorithms maintain good accuracy even a large proportion of the data is missing.

CONS

The following are the disadvantages of Random Forest algorithm –

- Complexity is the main disadvantage of Random forest algorithms.
- Construction of Random forests are much harder and time-consuming than decision trees.
- More computational resources are required to implement Random Forest algorithm.
- It is less intuitive in case when we have a large collection of decision trees.
- The prediction process using random forests is very time-consuming in comparison with other algorithms.

CHAPTER 4

RESULTS AND DISCUSSION

The question of which model type to apply to a Machine Learning task can be a daunting one given the immense number of algorithms available in the literature. It can be difficult to compare the relative merits of two methods, as one can outperform the other in a certain class of problems while consistently coming in behind for another class.

4.1 DECISION TREE BOUNDARIES

Logistic Regression and trees differ in the way that they generate decision boundaries i.e. the lines that are drawn to separate different classes. Decision Trees bisect the space into smaller and smaller regions, whereas Logistic Regression fits a single line to divide the space exactly into two. Of course for higher-dimensional data, these lines would generalize to planes and hyperplanes. A single linear boundary can sometimes be limiting for Logistic Regression. In this example where the two classes are separated by a decidedly non-linear boundary, we see that trees can better capture the Regression's simple linear boundary generalizes better.

Each node of a Decision Tree assigns a constant confidence value to the entire region that it spans, leading to a rather patchwork appearance of confidence values across the entire space. On the other hand, prediction confidence for Logistic Regression can be computed in closed-form for any arbitrary input coordinates, so that we have an infinitely more fine-grained result and can be more confident in our prediction confidence values.

division, leading to superior classification performance. However, when classes are not well-separated, trees are susceptible to overfitting the training data, so that Logistic

4.2 LOGISTIC REGRESSION

Logistic Regression is a performance advantage, but its resulting f-measure did not exactly beat the Decision Tree's by a huge margin. When a tree consists of a large number of nodes, it can require a significant amount of mental effort to comprehend all the splits that lead up to a particular prediction. In contrast, a Logistic Regression model is simply a list of coefficients. we are able to see that an coefficients to differ by

a factor of 1000 in order to represent equal importance in a real-world, physical sense. Because Logistic Regression models are fully described by their coefficients, they are attractive to users who have some familiarity with their data, and are interested in knowing the influence of particular input fields on the objective. determining its class, which is corroborated by the slope of the decision boundary from the instance's y-coordinate is just over three times as important as its x-coordinate for previous section. An important caveat to this is in regards to scale. If for example, x and y were given in units of meters and kilometres respectively, we should expect their output.

4.3 DATA PREPROCESSING

	0	1	2	3	4		5	6	7	8
0	989292962323615744	2018-04-25	23:59:57	Eastern Standard Time	whosalli	The lack of this understanding is a small but ...	1	0	3	
1	989292959844663296	2018-04-25	23:59:56	Eastern Standard Time	estermnunes	i just told my parents about my depression and...	1	0	2	
2	989292951716155392	2018-04-25	23:59:54	Eastern Standard Time	TheAlphaAries	depression is something i don't speak about ev...	0	0	0	
3	989292873664393218	2018-04-25	23:59:35	Eastern Standard Time	_ojhodgson	Made myself a tortilla filled with pb&j. My de...	1	0	0	
4	989292856119472128	2018-04-25	23:59:31	Eastern Standard Time	DMiller96371630	@WorldofOutlaws I am gonna need depression med...	0	0	0	

Fig.4.1: Preprocess Data

From Fig 4.1 it can require a significant amount of mental effort to comprehend all the splits that lead up to a particular prediction and to load data user able to give data.head.

4.4 TEXT TOKENIZATION

```
ain't": "am not",  
aren't": "are not",  
can't": "cannot",  
can't've": "cannot have",  
'cause": "because",  
could've": "could have",  
couldn't": "could not",  
couldn't've": "could not have",  
didn't": "did not",  
doesn't": "does not",  
don't": "do not",  
hadn't": "had not",  
hadn't've": "had not have",  
hasn't": "has not",  
haven't": "have not",  
he'd": "he would",  
he'd've": "he would have",  
he'll": "he will",  
he'll've": "he will have",  
he's": "he is",  
how'd": "how did",  
how'd'y": "how do you",  
how'll": "how will",  
how's": "how is",
```

Fig.4.2: Keyword Vectors

From fig 4.2 the keywords given to give the meaning to the short words that are in the data set and each word has individual frequency that frequency matches with the data set .

4.5 CLASSIFICATION

```
hist = model.fit(data_train, labels_train, \
                 validation_data=(data_val, labels_val), \
                 epochs=EPOCHS, batch_size=40, shuffle=True, \
                 callbacks=[early_stop])
```

Train on 8530 samples; validate on 2845 samples

Epoch 1/10
8530/8530 [-----] - 87s 10ms/step - loss: 0.1149 - accuracy: 0.9630 - val_loss: 0.0329 - val_accuracy: 0.9930

Epoch 2/10
8530/8530 [-----] - 84s 10ms/step - loss: 0.0345 - accuracy: 0.9923 - val_loss: 0.0339 - val_accuracy: 0.9919

Epoch 3/10
8530/8530 [-----] - 422s 49ms/step - loss: 0.0277 - accuracy: 0.9934 - val_loss: 0.0341 - val_accuracy: 0.9926

Epoch 4/10
8530/8530 [-----] - 79s 9ms/step - loss: 0.0293 - accuracy: 0.9936 - val_loss: 0.0409 - val_accuracy: 0.9902

Fig.4.3: Prediction

From fig 4.3 we train the data set and test it to get the accuracy of the given data in the given twitter data set and if you give new input it will predict based on training data

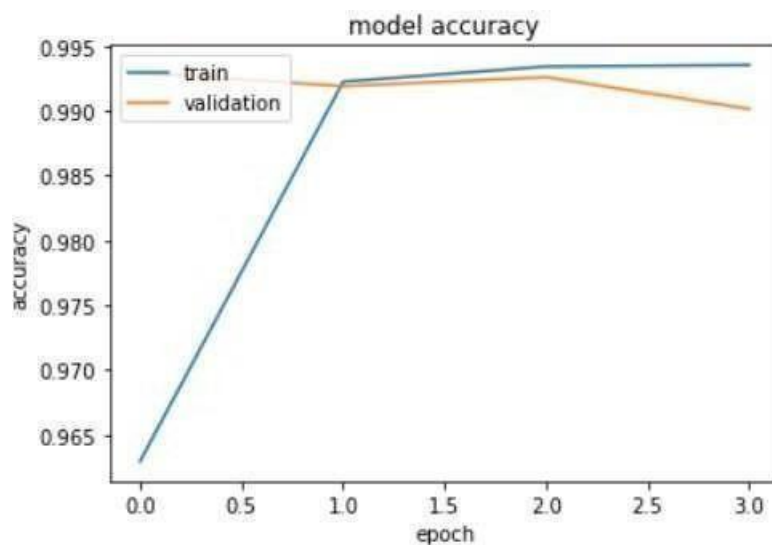


Fig.4.4 : Model Accuracy

From Fig 4.4 we get the model accuracy in the graph to give a perfect accuracy using the random forest algorithm based on training data user able to get data and predict the output.

CHAPTER 5

SUMMARY AND CONCLUSION

5.1 SUMMARY

This work is a proof of concept study demonstrating that indicators of future mental health can be extracted from people's natural language using computational methods. Our approach included the development of a mathematical algorithm for unpacking the meaning components of a sentence as well as a computational pipeline for identifying the kinds of thought content that are potentially diagnostic of mental illness. Finally, we showed how the linguistic indicators of mental health, semantic density and talk about voices, could predict the onset of psychosis at high levels of accuracy.

In addition to predicting the onset of psychosis, the methods provide insight into the thought processes affected in the emergence of psychosis. Our results from randomly shuffling words showed that semantic density was a function of the way words were organized into sentences, not simply which words were used across sentences. The results are thus consistent with work suggesting that patients with psychosis have impairments in the integration of words to generate higher order meaning.³⁷ These results thus extend the literature on how computational methods can be used to predict and diagnose psychosis from speech. In an analysis of the speech, Mota et al.¹³ could distinguish those with mania from schizophrenia by drawing on a set of ten different graph measures with 93.7% accuracy.¹³ Also based on the graph analyses of thought disorganization, Mota et al.¹² found that connectedness of language was lower in schizophrenia patients than in healthy control participants 6 months in advance of the schizophrenia patients receiving a diagnosis, allowing prediction accuracy of 91.7%.¹² Elvevåg B, et al.⁹ differentiated schizophrenia from healthy controls using semantic coherence based on average cosine similarity of words.⁹ This method was extended by Bedi et al.,¹⁰ who found that the combination of semantic coherence, the number of determiner pronouns and maximum phrase length, could predict conversion to schizophrenia in the training set

with 100% accuracy. Lastly, Corcoran et al.¹¹ found that four predictor variables in free speech—maximum coherence, variance coherence, minimum coherence, and possessive pronouns—could be used to predict the onset of psychosis with 83% accuracy. In addition to measuring abnormal thought processes, the current study offers a method for the early detection of abnormal auditory experiences at a time when such abnormalities are likely to be missed by clinicians.

5.2 CONCLUSION

With the exception of Alzheimer's dementia, schizophrenia is the first mental disorder to which the prediction and prevention program of modern medicine has hitherto systematically been applied. The results are promising and justify the expectation that in the years to come it will be possible to provide preventive strategies tailored specifically to the individual risk of illness of each person seeking advice. In order to attain a major reduction in incidence, symptom-oriented risk assessment has to be enriched by neurobiological and psychosocial risk factors, and indicated prevention has to be further developed towards selective prevention. This requires a new generation of large sample studies for prediction as well as prevention, with significantly longer observation periods. In these studies, promising combinations of risk indicators, selected to maximize predictive values, must be evaluated, psychological and pharmacological interventions need to be assessed on a long-term basis, more etiologically oriented prevention strategies have to be tested. In order to be able to plan and conduct such studies, it would certainly be helpful to include subpsychotic mental states, as defined by the currently used risk symptoms, in the upcoming revision of the diagnostic systems.

5.3 FUTURE WORK

In this novel system, we implemented an artificial intelligence application to division normal and abnormal Schizophrenia subjects from social media networks using machine learning methodology. This work proposes a LSTM+CNN based framework for automated Schizophrenia Discourse detection, to possibly gather connection phrases from tweets text contents. The semantic consistence between the machine and human affiliation systems is tentatively confirmed. This discovering gives

a knowledge of comprehension the affiliated. From the all statements described in this paper, we can easily conclude that all the phases engaged in this approach are important and they all impact to obtaining the finest out of the algorithms. This complete procedure stream developed in additional precise analysis prediction of Schizophrenia discourse. In the future, we would like to extend this project by analysing for other diseases in the same mental illness areas.

REFERENCES

- [1] Allen.E, Damaraju,.D Plis.S, Erhardt.E, Eichele.T, Calhoun.V.D, -Tracking wholebrain connectivity dynamics in the resting state,| Cereb. Cortex, vol.24(3), 2013, pp.663-676, 2013.
- [2] Agirre Eneko, Bengoetxea Kepa, Joakim Nivre, Yue Zhang, and Koldo Gojenola. 2014. On wordnet semantic classes and dependency parsing. In Proceedings of the 52th Annual Meeting of the Association of Computational Linguistics pages 649–655, Baltimore (Maryland) - (USA), June. Association for Computational Linguistics.
- [3] Bansala Mohit, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22 -27, 2014, Baltimore, MD, USA, Volume 2: Short Papers, pages 809–815.
- [4] Bengoetxea Kepa, Eneko Agirre, Joakim Nivre, Yue Zhang, and Koldo Gojenola. 2014. On wordnet semantic classes and dependency parsing. pages 649– 655. ACL 2014.
- [5] Iakes Goenaga, Koldo Gojenola, and Nerea Ezeiza. 2014. Combining clustering approaches for semisupervised parsing: the basque team system in the sprml2014 shared task. In First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages: Shared Task on Statistical Parsing of Morphologically Rich Languages, Dublin, Ireland, August. Dublin City University.
- [6] Patel. P, Aggarwal .P, and Gupta .A, -Classification of schizophrenia versus normal subjects using deep learning,| in Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, ser. ICVGIP '16. New York, NY, USA: ACM, 2016, pp. 28:1–28:6. [Online]. Available: <http://doi.acm.org/10.1145/3009977.30>

- [8] Project Information 12 4 Johannesen J.K, J. Bi, Jiang .R, Kenney J.G, and C. A. Chen, -Machine learning identification of eeg features predicting working memory performance in schizophrenia and healthy adults,| Neuropsychiatric Electrophysiology, vol. 2, no. 1, p. 3, 2016.
- [9] Seixas F.L, Zadrozny .B, Laks.B, Conci .A, and D. C. Muchaluat Saade, -A bayesian network decision model for supporting the diagnosis of dementia, alzheimer's disease and mild cognitive impairment,| Computers in Biology and Medicine, vol. 51, pp. 140–158, 2017.
- [10] Seddah.D, Sandra K ´ubler, and Reut Tsarfaty. " 2014. Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. In Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages, pages 103–109, Dublin, Ireland, August. Dublin City University.
- [11] Song .H, Chen.G, Gao .R, I. Bogdan. I, Yang.J, Wang.S, Dong.W, Quan,.W Dang.W, and X. Yu, -Automatic schizophrenic discrimination on fnirs by using complex brain network analysis and svm,| BMC Medical Informatics and Decision Making, vol. 17, no. 3, p. 166, Dec 2017.
- [12] Rongjun Yu, Yi-Ling Chien, Hsiao-Lan Sharon Wang, Chih-Min Liu, Chen-Chung Liu, Tzung-Jeng Hwang, H. Ming Hsieh, Hai-Gwo Hwu, and Wen-Yih Isaac Tseng, -Frequency-Specific Alternations in the Amplitude of Low-Frequency Fluctuations in Schizophrenia,| Human Brain Mapping, vol.35, 2014, pp.627-637.

APPENDIX

A.SOURCE CODE

```
#!/usr/bin/env python

# coding: utf-8

import warnings

warnings.filterwarnings("ignore")

import numpy as nP

import pandas as pd

import matplotlib.pyplot as plt

import nltk

import ftfy

import re

from math import exp

from numpy import sign

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

from gensim.models import KeyedVectors

from nltk.corpus import stopwords

from nltk import PorterStemmer

from keras.models import Model, Sequential

from keras.callbacks import EarlyStopping, ModelCheckpoint

from keras.layers import Conv1D, Dense, Input, LSTM, Embedding, Dropout,
Activation, MaxPooling

from keras.preprocessing.text import Tokenizer
```

```

from keras.preprocessing.sequence import pad_sequences

np.random.seed(1234)

EPOCHS= 10

LEARNING_RATE = 0.1

TEST_SPLIT = 0.2

DEPRES_NROWS = 3200

RANDOM_NROWS = 12000

MAX_SEQUENCE_LENGTH = 140


MAX_NB_WORDS = 20000

EMBEDDING_DIM = 300

TRAIN_SPLIT = 0.6

DEPRESSIVE_TWEETS_CSV = 'SchizophrenicDiscourseDatasets.csv'

RANDOM_TWEETS_CSV = 'NonSchizophreniaDataset.csv'

EMBEDDING_FILE = 'GoogleNews-vectors-negative300.bin.gz'

random_tweets_df = pd.read_csv(RANDOM_TWEETS_CSV, encoding = "ISO-8859-1", usecols = range(0,4), nrows = RANDOM_NROWS)

depressive_tweets_df.head()

random_tweets_df.head()

word2vec = KeyedVectors.load_word2vec_format(EMBEDDING_FILE, binary=True)


depressive_tweets_df = pd.read_csv(DEPRESSIVE_TWEETS_CSV, sep = '|', header = None, usecols = range(0,9), nrows = DEPRES_NROWS)

```

```
cList = {  
  
    "ain't": "am not",  
  
    "aren't": "are not",  
  
    "can't": "cannot",  
  
    "can't've": "cannot have",  
  
    "'cause": "because",  
  
    "could've": "could have",  
  
    "couldn't": "could not",  
  
    "couldn't've": "could not have",  
  
    "didn't": "did not",  
  
    "doesn't": "does not",  
  
    "don't": "do not",  
  
    "hadn't": "had not",  
  
    "hadn't've": "had not have",  
  
    "hasn't": "has not",  
  
    "haven't": "have not",  
  
    "he'd": "he would",  
  
    "he'd've": "he would have",  
  
    "he'll": "he will",  
  
    "he'll've": "he will have",  
  
    "he's": "he is",  
  
    "how'd": "how did",  
  
    "how'd'y": "how do you",)}}}
```

```
c_re = re.compile('%s' % '|'.join(cList.keys()))
```

```
c_re
```

```
def expandContractions(text, c_re=c_re):
```

```
    def replace(match):
```

```
        return cList[match.group(0)]
```

```
    return c_re.sub(replace, text)
```

```
def clean_tweets(tweets):
```

```
    cleaned_tweets = []
```

```
    for tweet in tweets:
```

```
        tweet = str(tweet)
```

```
        if re.match("(\\w+:\\V\\S+)", tweet) == None and len(tweet) > 10
```

```
            tweet = ' '.join(re.sub("(@[A-Za-z0-9]+)|(#[A-Za-z0-9-9]+)|(<Emoji:.*>)|(pic\\.twitter\\.com\\V\\.*)", " ", tweet).split())
```

```
            tweet = ftfy.fix_text(tweet)
```

```
            tweet = expandContractions(tweet
```

```
            tweet = ' '.join(re.sub("([0-9A-Za-z \\t])", " ", tweet).split(
```

```

stop_words = set(stopwords.words('english'))

word_tokens = nltk.word_tokenize(tweet)

filtered_sentence = [w for w in word_tokens if not w in stop_words]
tweet = ' '.join(filtered_sentence)

tweet = PorterStemmer().stem(tweet)

cleaned_tweets.append(tweet)

return cleaned_tweets

depressive_tweets_arr = [x for x in depressive_tweets_df[5]]

depressive_tweets_arr

random_tweets_arr = [x for x in random_tweets_df['SentimentText']]

X_d = clean_tweets(depressive_tweets_arr)

X_d

X_r = clean_tweets(random_tweets_arr)

tokenizer = Tokenizer(num_words=MAX_NB_WORDS)

tokenizer.fit_on_texts(X_d + X_r)

sequences_d = tokenizer.texts_to_sequences(X_d)

sequences_d

sequences_r = tokenizer.texts_to_sequences(X_r)

sequences_r

```

```

word_index = tokenizer.word_index

print('Found %s unique tokens' % len(word_index))

data_d = pad_sequences(sequences_d, maxlen=MAX_SEQUENCE_LENGTH)

data_r = pad_sequences(sequences_r, maxlen=MAX_SEQUENCE_LENGTH)

print('Shape of data_d tensor:', data_d.shape)

print('Shape of data_r tensor:', data_r.shape)

nb_words = min(MAX_NB_WORDS, len(word_index))

embedding_matrix = np.zeros((nb_words, EMBEDDING_DIM))

for (word, idx) in word_index.items():

    if word in word2vec.vocab and idx < MAX_NB_WORDS:

        embedding_matrix[idx] = word2vec.word_vec(word)

labels_d = np.array([1] * DEPRES_NROWS)

labels_r = np.array([0] * RANDOM_NROWS)

perm_d = np.random.permutation(len(data_d))

idx_train_d = perm_d[:int(len(data_d)*(TRAIN_SPLIT))]

```

```

idx_test_d =
perm_d[int(len(data_d)*(TRAIN_SPLIT)):int(len(data_d)*(TRAIN_SPLIT+TEST_SPLIT))]

idx_val_d = perm_d[int(len(data_d)*(TRAIN_SPLIT+TEST_SPLIT)):]

perm_r = np.random.permutation(len(data_r))

idx_train_r = perm_r[:int(len(data_r)*(TRAIN_SPLIT))]

idx_test_r =
perm_r[int(len(data_r)*(TRAIN_SPLIT)):int(len(data_r)*(TRAIN_SPLIT+TEST_SPLIT))]

idx_val_r = perm_r[int(len(data_r)*(TRAIN_SPLIT+TEST_SPLIT)):]

data_train = np.concatenate((data_d[idx_train_d], data_r[idx_train_r]))

labels_train = np.concatenate((labels_d[idx_train_d], labels_r[idx_train_r]))

data_test = np.concatenate((data_d[idx_test_d], data_r[idx_test_r]))

labels_test = np.concatenate((labels_d[idx_test_d], labels_r[idx_test_r]))

data_val = np.concatenate((data_d[idx_val_d], data_r[idx_val_r]))

labels_val = np.concatenate((labels_d[idx_val_d], labels_r[idx_val_r]))

perm_train = np.random.permutation(len(data_train))

data_train = data_train[perm_train]

labels_train = labels_train[perm_train]

```

A Unique & Dynamic Approach to Predict Schizophrenia Disease using Machine Learning

Detection of Schizophrenia Discourse from Twitter Tweets

Neelisetti Ashok
CSE Department
Sathyabama University
Chennai, India
neelisettiashok@gmail.com

Tatikonda venkata sai manoj
CSE Department
Sathyabama University
Chennai, India
saimanoj1010@gmail.com

D.Usha Nandini
CSE Department
Sathyabama University
Chennai, India
usha.cse@sathyabama.ac.in

Abstract— The society environment and its wide array of influences factors are very much connected with each phase of human life operating and development. Schizophrenia is a chronic and severe mental disease which alters the human memory and mind. Mental fitness and mental disease are dictated by various and collaborating social, mental, what's more, organic variables, similarly as wellbeing and sickness when all is said in done. The clearest proof identifies with the dangers of mental ailments, which in the created and creating world are related with markers of neediness, including low degrees of training. The singular goal of this proposed project is to automatically distinguish patients with schizophrenia and without schizophrenia. Detection of schizophrenia in initial periods is crucial to prevent comprehensive mental illness. Currently, identifying the schizophrenia discourse which are extensively distributed in our culture and forecasting the forthcoming phases of the mental illness has become extremely crucial in modern-day society. Normally patients express their health condition in one of the most popular social media networks called Twitter. We introduce a unique and dynamic machine learning (ML) approach to predict the schizophrenia discourse from twitter tweets. This unique approach is extremely adaptable as practically any supervised machine learning algorithm can be utilized for prediction. The efficiency and accuracy of the machine learning algorithm is directly proportional to the excellence of the feature engineering which needs a high level of domain knowledge. To overcome this problem, this paper proposes a Deep Neural Network (DNN) which is capable of automatically identifies the pattern of schizophrenia discourse and classifies the schizophrenia discourse.

Keywords— *Machine Learning, Deep Learning, Convolutional Neural Network CNN, Long-Short Term Memory LSTM, Natural Language Processing, Discourse Comprehension, Schizophrenia*

I. INTRODUCTION

Schizophrenia is a dangerous neurological sickness, which seems in roughly 1% of the adult population around the globe. It affects remembrance, awareness, linguistic and executive functions of the brain. According to World Health Organization (WHO) stated that the worldwide incidence of global Schizophrenia is just about 20 million people. Among

that 69% percentage of Schizophrenia disorder patients are not obtaining proper meditations.

There are various causes for Schizophrenia disorders. Any research does not distinguish one particular aspect. Hereditary and environmental influences perform a major role in its diagnosis. It is normally developed in the adulthood under the age of 26. The patients which are affected by this disease will have intellectual deficiencies, including reading disorders. They will have difficulty to select suitable phrases while they are talking and composing. It modifies patients' observation, points of view, and conduct as confirm by fantasies, dreams, muddled discourse or conduct, social withdrawal, and differed intellectual deficiencies.

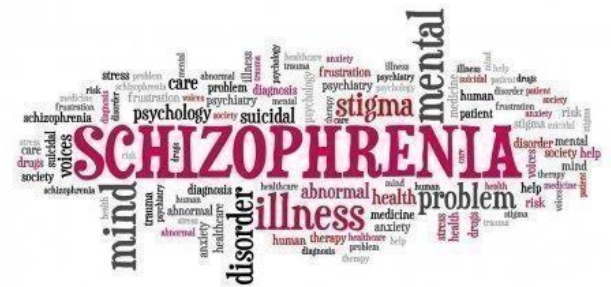


Fig 1. Schizophrenia Word Cloud

Schizophrenia Discourse (SD) is a mental ailment, which is the highly predominant neurodegenerative disease in this globe. It is a radical and dense mental illness influencing susceptible peoples, whose pervasiveness is projected to twice by 2030. Diagnosis and medication of neurological Discourse are emphases in brain study. Despite the fact that meds have been created to treat schizophrenia, they don't work for all patients, and there is proof that a few types of schizophrenia seem, by all accounts, to be treatment safe. This just underlines the significance of leading fruitful schizophrenia clinical preliminaries.

According to the artificial intelligence study, social media text analysis has been a leading area of study in the past two to three decades that has been commonly applied in wide array

of research such as healthcare, smart home, surveillance systems, human-computer interaction, gaming and so on.

With an exceptional development of web-based social networking, huge number of individuals deliberately share a lot of information by communicating their states of mind, sentiments, feelings, and day by day battles with emotional well-being issues via web-based networking media stages like Twitter. This offers open doors for new comprehension of these networks. With the improvement of Internet utilization, individuals have begun to impart their encounters and difficulties to emotional wellness issue through online discussions, small scale web journals or tweets. Their online exercises roused numerous specialists to present new types of potential human services arrangements and techniques for early melancholy recognition frameworks. Utilizing distinctive Natural Language Processing (NLP) methods and content characterization draws near, they attempted to prevail in a better improvement.

The singular aim of this application is to develop a machine learning especially deep learning approach, named Long-Short Term Memory (LSTM) plus Convolutional Neural Network (CNN), for detecting Schizophrenia disorders using features obtained from a different dataset from kaggle.com, namely the Schizophrenia Discourse dataset. The proposed LSTM+CNN is competent to handle high-ranking dimensional data by expanding the model interpretability and discovering the highly appropriate features while dispensing with the typical inequitable class division.

Natural language processing (NLP) has gained significantly from the revival of neural networks (NNs), expected to their superior execution with fewer necessary of contrived features. Convolutional neural network (CNN) and recurrent neural network (RNN) are the two major types Deep neural network architecture. Gating systems have been established to improve certain constraints of the basic recurrent neural network, causing in two fundamental recurrent neural network types: long short-term memory (LSTM) and gated recurrent unit (GRU).

This application is therefore dedicated on approaches to categorize symptoms of Schizophrenia Discourse in posted tweets as early as feasible established on deep learning methodology. This paper content is organized into six divisions. The first division contains the Introduction section. The second division specifies about social network. The third division illuminates related work. Our proposed system is deliberated in the fourth division which covers the datasets, data analysis procedure, the proposed LSTM+CNN approach. The experiment results are presented in the fifth division and conclusions are illustrated in the final segment. The final segment also contains the discussion as well as the future work in that direction.

II. SOCIAL NETWORK

The social media network has become immeasurable and reputation along with all the service industries today. Today most of the people around the world are using the social media network to share their individual, qualified interest and condition consciousness.

Social media networks are the collection of collaborative web-based technology where interested communities produce and share expertise, concepts, intelligence, specialized interests and other information through virtual communities using online communication channels. There are various types of social medias like wikis, Social Networks, Media Sharing Networks, Discussion Forums, Bookmarking & Content Curation Networks, Social Review Websites, Community Blogging & Publishing Websites, Image & Video Hosting Networks, Sharing Economy Networks, Social Shopping Networks, Interest-Based Networks, Anonymous Social Networks and many more.

The popular social networks are Facebook, Twitter, LinkedIn etc. The social networks are web based social media applications which are constantly evolving. The communities uses the social network to connect people and share knowledge through human-to-human interaction. It is also called as relationship networks where like minded people can share ideas and knowledge.

The popular image and video hosting websites are Flickr, photobucket, Instagram, Snapchat, YouTube, Vimeo etc. It is also called as Media Sharing Websites. These media sharing networks can be used to share photos, images, videos, live videos, audios and other medias which can be accessed from anywhere and anytime around the world. The image and video hosting websites can be used to promote brand recognition, lead creation etc. These networks mostly promote user generated content.

The widespread Discussion Forums networks are reddit, Quora, Digg, Warrior Forum, etc. Discussion forum are intrinsically social; since it brings community people collectively on a collective ground. In these networks, people participate in discussions by publishing and replying to community messages and also share news, information and other options. The discussion forum can be either public or private depending on the conversation intentions. These networks are highly effective for the market research team.

The predominant social review networks are TripAdvisor, Yelp, FourSquare, Google My Business, BBB (Better Business Bureau) etc. These networks allow the online visitors to determine, share and review for all sorts of location and experiences like services, products and brands as well as restaurants, travel destinations, and just about anything else. These networks are highly effective to understand the

customer's and consumer's perspectives. It also provides a valuable feedback for any kind businesses.

The common Community Blogging & Publishing Websites are Medium, Tumblr, Blogger, Wordpress etc. A blogging platform is a web based application which can be used to publish your creative content online. These network can be used to publish and find the online content. It is the most successful methodology to engage with your customers or audience. Few Blogging networks are exclusive. This means that it can be only access by invite.

The widespread Sharing Economy Networks are Airbnb, Pantheon, Kickstarter etc. These kind networks can be used to promote, discover, distribute, purchase, trade and export & import products and services between community. It is also called as peer economy or gig economy network. Few online marketplace offers tourism experiences.

The predominant Bookmarking & Content Curation Networks are Pinterest, Flipboard, Digs, StumbleUpon etc. The online visitors can use these Bookmarking & Content Curation Networks to save and organize hyperlinks to any number of web resources and web pages. These applications allow the users to categorize the web links which can be shared among their followers. These categories and web links can be easily searchable. These networks can be used to discuss new and trending messages, images and videos. The primary focus on Content Curation Networks are to determine and share trending articles and other text content.

III. RELATED WORK

Most of the existing Natural Language Processing framework such as Stanford CoreNLP and SpaCy are only concentrates on predicting semantic forms rather than modeling Natural Language Processing (NLP) designs. Many of the recent research studies have explored linguistics expressions of Schizophrenia Disease in social network dataset especially from Twitter. Some of the recent research studies focuses on mental illness and their moods and some of the other research papers focuses on Schizophrenia Disease classifications. Attention Discrepancy Hyperactivity Syndrome [2] classifies the patients correlated with chronic Schizophrenia Disease. In this research, the dataset is extracted from the Reddit which is a famous social network community platform. This paper uses the Reddit API to develop the corpus. It implemented using the co-training methodology with the widely used classifier algorithms like Support Vector Machine, and Naïve Bayes.

In the study [4] detects the mental health expression and where the corpus is extracted from college student influenced data from Reddit API. This paper analysis the verbal and chronological attributes of the Schizophrenia Diseases from college students in online social media communities. It implements a novel inductive transfer learning approach. It

applies the binary machine learning classification algorithm with positive and negative as class labels. The machine learning classification algorithm like logistic regression, random forest, Ada Boost etc are used for classifying the Schizophrenia Diseases.

In research study [9], inventors used Deep Beliefs Networks (DBN) which adopts a layer by layer training approach. The dataset contains high thin dimensionality feature vectors. Deep Beliefs Networks (DBN) is used solving this high dimensionality problem and also decreases the computational complexity from extracting the high level feature to low level features. The classification of the data is implemented using multinomial logistic regression. multinomial logistic regression is a multi-class classification framework. The outcome of this study is better than Support Vector Machine and Neural Network. In multinomial logistic regression, the sigmoid logistic function is replaced by SoftMax function ϕ . This research defines the net input Z is

$$Z = w_0x_0 + w_1x_1 + \dots + w_mx_m = \sum_{i=0}^m w_i x_i = w^T x$$

In this research [10], authors performs the automatic text classification from the medical documents. Dataset are gathered from medical hospital which contains the information about 8 different reasons of death. In the preprocessing steps, authors change the corpus to the lower case statements. They have also removed the special characters, punctuations and stop words using the Natural Language Processing Toolkit (NLTK). After that, unigram, bigram, trigram, and hybrid-gram features were obtained from the prep-processed corpus. In this manner, creators have utilized distinctive best in class content orders models, for example, SVM, NN, NB, and Random Forest (RF). At long last, creators detailed the SVM as an outflanked classifier when contrasted with others.

In study [12], creators have performed a research to evaluates the accomplishment of wide range of techniques for automatic conceptual overview of text classification. We recognize opinion and substance grouping, which are both quite compelling for promoting applications. In slant order, the objective of the content grouping strategies is distinguish the feeling passed on through this content and to effectively order it as negative. Dictionary based or regulated AI strategies are the two significant ways to deal with achieve this errand. The author has used five machine learning algorithms 1. Artificial Neural Network (ANN) 2. K-Nearest Neighbours (KNN) 3. Naïve Bayes (NB) 4. Random Forest (RF) and 5. Support Vector Machine (SVM). The authors used five lexicon based methods such as 1. AFINN-en-165.txt 2. Bing Liu and collaborators 3. nrc from Saif Mohammad and Peter Turney 4. ADER (Valence Aware Dictionary for sEntiment

Reasoning). The first three of these lexicons are based on unigrams, i.e., single words.

Given the above investigations, we can reason that SVM, NB and Random Forest perform better for classification of content. It fundamentally relies upon the attributes of a Dataset which is under investigation. The vast majority of the investigations utilized basic SVM or SVM with some alteration so as to show signs of improvement classification accuracy. Likewise, for RF it tends to be deduced that it is generally excellent for multi-class classification and NB additionally functions admirably with the dataset of little example sizes. Subsequently, the Machine learning approaches beat the dictionary based calculations. Besides, it tends to be seen that feature extraction and feature determination procedures can help improve the classification accuracy.

IV. PROPOSED WORK

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Data Acquisition

To train the machine learning algorithm, we require two different types of datasets: 1. Schizophrenia discourse dataset 2. Non-Schizophrenia discourse dataset. One of the major steps for neuroscience is Data gathering or Data Acquisition; since a minor error in the data acquisition may possibly trigger imprecision in prediction analysis after the predictive model is built. Twitter is a web based micro blogging and social networking service provider in which online visitors can post and obtain messages which can be in the maximum size of 140 characters. In Twitter, these messages are referred as tweets. The Kaggle.com already collected the tweets related to Schizophrenia discourse and provides us to download those tweets as datasets. To acquire these clients, Kaggle.com arbitrarily inspected Twitter IDs. From that point, we continued to download the tweets from the chose IDs .

Sample Schizophrenia Dataset

2018-04-25	07:01:08	Eastern Standard Time	WhatTheFFacts	Many sufferers of depression aren't sad, they ...
2018-04-25	07:01:01	Eastern Standard Time	luvsguap	No Depression by G Herbo is my mood from now o...
2018-04-25	07:00:48	Eastern Standard Time	AyJaxxWhitney	What do you do when depression succumbs the br...
2018-04-25	07:00:48	Eastern Standard Time	joseholist	Ketamine Nasal Spray Shows Promise Against Dep...
2018-04-25	07:00:42	Eastern Standard Time	CampfireLucifer	dont mistake a bad day with depression! everyo...

Sample Non-Schizophrenia Dataset

ItemID	Sentiment	SentimentSource	SentimentText
0	1	0	Sentiment140 is so sad for my APL frie...
1	2	0	Sentiment140 I missed the New Moon trail...
2	3	1	Sentiment140 omg its already 7:30 :O
3	4	0	Sentiment140 .. Omgaga. Im sooo im gunna CRy. I'...
4	5	0	Sentiment140 i think mi bf is cheating on me!!! ...

Tweets Summary

	Schizophrenia Tweets	Normal Tweets
Training	1384	7146
Validation	462	2382
Testing	462	2383
Total	2308	11911

B. Preprocessing

This research only contemplates the English authored subject. The next phase of the proposed approach is data preprocessing. In any machine learning algorithms, the preprocessing of the dataset normally contains various steps.

Step 1: Rephrase all tweeter seeds as phrases or unigrams.

Step 2: Delete all retweet handlers

Step 3: Delete all the URL from each tweets

Step 4: Delete all Non-ASCII characters and words from the Schizophrenia discourse dataset and Non-Schizophrenia discourse dataset

Step 5: Break Usernames and Hashtags

Step 6: Delete any punctuations from the hashtags.

Step 7: Delete all punctuations from the tweets

Step 8 : Load the prebuilt datasets and remove those stop words from datasets

Step 9: Delete all leading and trailing spaces.

Step 10: Delete all numbers from the tweets

Step 11: Delete all multiple in between text white spaces from the tweets

C. Natural language Processing (NLP)

Natural language processing provides efficient methodology to investigate the text content. Inheritance Structures and Significance of the text content can be determined using the Syntax analysis and Semantic analysis. One of the major tasks for NLP is to classify the sequence. In other words, it is used to designate the categorical label for each word in the text content. This can be implemented using

- 1) Part of Speech Tagging
- 2) Named Entity Recognition

Part of Speech Tagging

Parts of Speech (POS) Tagging can be used for information retrieval. There are two main kinds of tagsets available in Natural Language Processing

1. Brown Corpus Tagset
2. Penn Treebank Tagset

Few of the BrownCorpusTagset

Tag	Description
CC	Coordinating Conjunction
CD	Cardinal Numeral
CS	Subordinating Conjunction
DT	singular determiner
DTI	singular or plural determiner/quantifier
DTS	plural determiner
DTX	determiner/double conjunction
FW	foreign word
HL	word occurring in headline
IN	Preposition
JJ	Adjective
JJR	Comparative Adjective
JJS	Semantically Superlative Adjective
JJT	Morphologically Superlative Adjective
NN	Singular Or Mass Noun

Few of the Penn Treebank tagset

Tag	Description
CC	Coordinating Conjunction
CD	Cardinal Numeral
EX	Existential There
IN	Preposition, Subordinating Conjunction
POS	Possessive Ending
PP	Personal Pronoun
PP\$	Possessive Pronoun
RB	Adverb
RBR	Adverb, Comparative
RBS	Adverb, Superlative
SENT	Sentence-Break Punctuation
VB	Verb Be, Base Form

POS labeling is one of the apparatuses and the segments for foundation to Natural Language Processing of a given language. POS labeling is vital in numerous fields, for example, content expression, language structure, semantic investigation and interpretation

A tagset is a vector of parts of speech tags. It designates every single word the extremely expected within the framework. There are three methodologies are available for Parts of Speech (POS) Tagging.

- 1) Rule-Based POS tagging
- 2) Transformation-based tagging
- 3) Stochastic (Probabilistic)tagging.

Let us consider

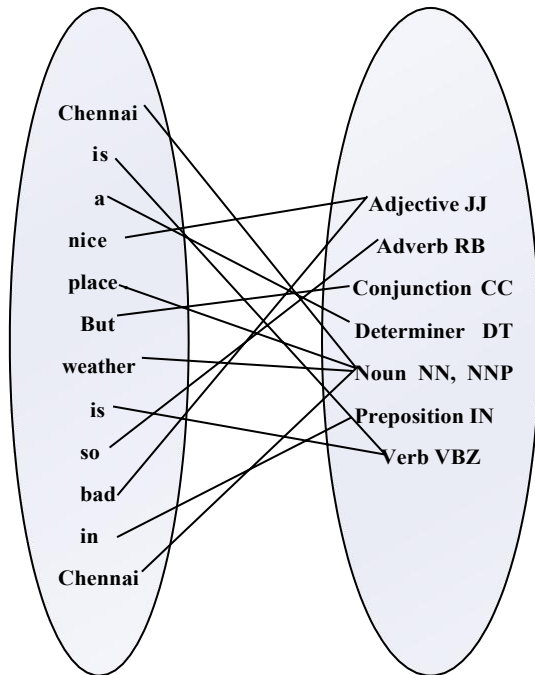
$W = w_1 w_2 w_3 \dots w_n$ Series of words

$T = t_1 t_2 t_3 \dots t_n$ Series of POS Tags

$$f : W \rightarrow T = f(W)$$

Rule based POS Tagging

Calculation for allocating grammatical feature dependent on two phases, the main stage utilizes a vocabulary (lexicon) to dole out each word a rundown of potential grammatical features, the subsequent stage utilizes an enormous rundown of manually written disambiguation rules to winnow down this rundown to a solitary grammatical feature for each word

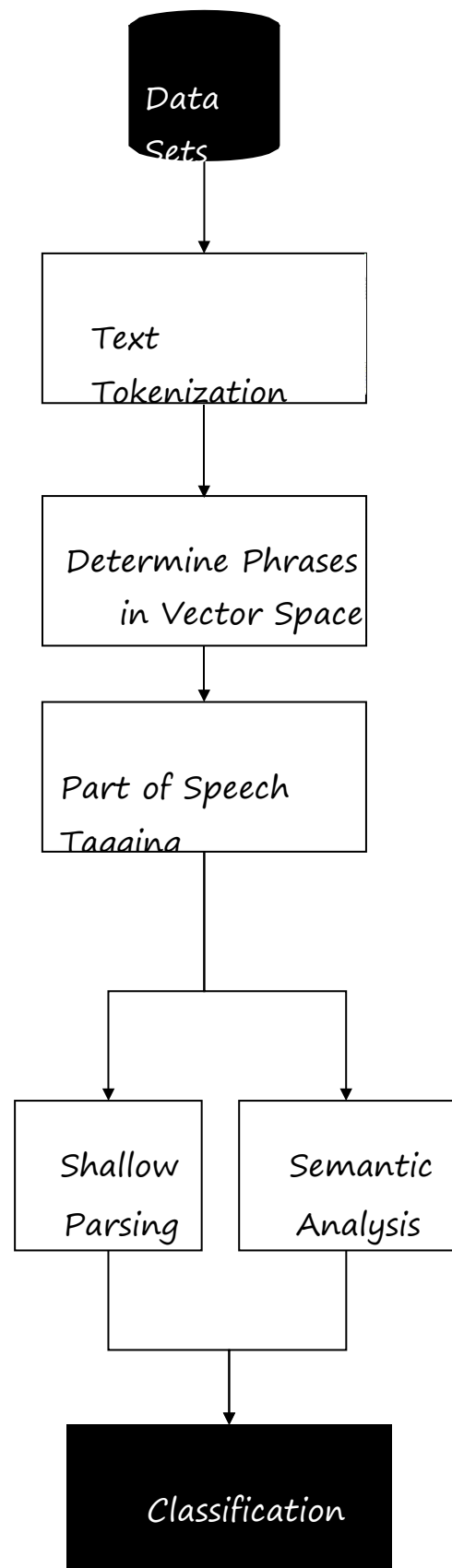


Named Entity Recognition

Named Entity Recognition (NER) can perform the process of entity identification and information retrieval. It is capable of classifying the named entities which are present in the text content into pre-defined categories like people, locations, organizations, percentages, quantities, countries, dates, individuals, product information, service information etc. It provides a solid semantic meaning for the text content so that it is easily to understand the subject of any text content.

Named Entity Recognition (NER) is a sub issue of data e xtraction and includes handling organized and unstructured records and distinguishing articulations that allude to people groups, spots, associations and organizations. NER is an essential errand and it is the center of regular language handling (NLP) framework. NER includes two assignments, which are: right off the bat the recognizable proof of appropriate names in content, and also the classification of these names into a lot of predefined classes of intrigue, for e xa mple, persons, organizations, areas, date and time articulations

D. The Proposed Architecture



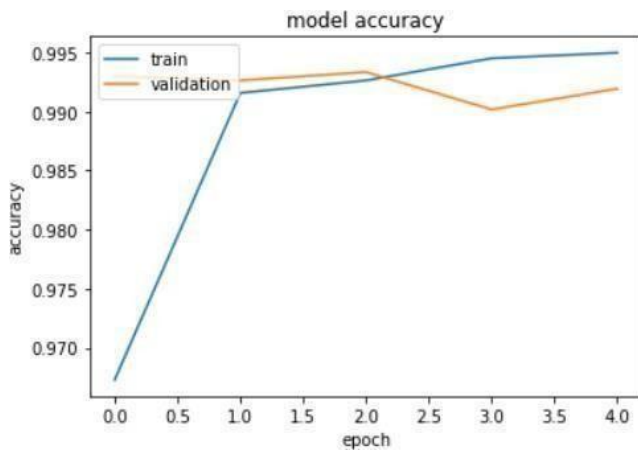
V. EXPERIMENT RESULTS

Performance Evaluation

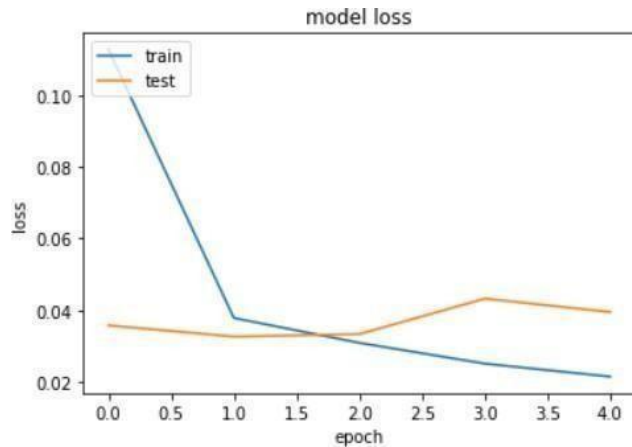
Class	Performance Measures		
	Precision	Recall	F-measure
Anxiety	0.89	0.80	0.82
ADHD	0.66	0.67	0.62
Depression	0.84	0.87	0.81
Bipolar	0.60	0.80	0.70

From the above table, we can monitor the efficiency of suggested methodology in terms of class wise categorization of tweeter statements. In terms of F-measures, we monitor the accomplishment of recommended method for the categorization of Anxiety tweets and its related comments (F-measure = 0.82) as compared to ADHD (F-measure = 0.62), Depression (F-measure = 0.84), and Bipolar (F-measure = 0.70).

Summarize history for accuracy



Summarize history for loss



VI. CONSLUSION & FUTURE WORK

In this novel system, we implemented an artificial intelligence application to division normal and abnormal Schizophrenia subjects from social media networks using machine learning methodology. This work proposes a LSTM+CNN based framework for automated Schizophrenia Discourse detection, to possibly gather connection phrases from tweets te xt contents. The semantic consistence between the machine and human affiliation systems is tentatively confirmed. This discovering gives a knowledge of comprehension the affiliated. From the all statements described in this paper, we can easily conclude that all the phases engaged in this approach are important and they all impact to obtaining the finest out of the algorithms. This complete procedure stream developed in additional precise analysis prediction of Schizophrenia discourse.

In the future, we would like to e xtend this project by analysing for other diseases in the same metal illness areas.

REFERENCES

- [1] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. Muchaluat Saade, "A bayesian network decision model for supporting the diagnosis of dementia, alzheimer׳s disease and mild cognitive impairment," *Computers in Biology and Medicine*, vol. 51, pp. 140–158, 2017.
- [2] Subhan Tariq, Nadeem Akhtar, Humaira Afzal, Shahzad Khalid, Muhammad Rafiq Mufti , Shahid Hussain , Asad Habib , Ghufuran Ahmad. A novel Co-training based approach for the classification of mental illnesses using Social media posts 2019
- [3] Rongjun Yu, Yi-Ling Chien, Hsiao-Lan Sharon Wang, Chih-Min Liu, Chen-Chung Liu, Tzung-Jeng Hwang, H. Ming Hsieh, Hai-Gwo Hwu, and Wen-Yih Isaac Tseng, " Frequency-Specific Alternations in the Amplitude of Low-Frequency Fluctuations in Schizophrenia," *Human Brain Mapping*, vol.35, 2014, pp.627-637.
- [4] S. Bagroy, P. Kumaraguru, and M. De Choudhury, " A Social Media Based Index of Mental Well-Being in College Campuses," pp. 1634–1646, 2017.
- [5] E. Allen, E. Damaraju, S.M. Plis, E. Erhardt, T. Eichele, V.D. Calhoun, "Tracking whole-brain connectivity dynamics in the resting state," *Cereb. Cortex*, vol.24(3), 2013, pp.663-676,2013.
- [6] Eneko Agirre, Kepa Bengoetxea, Joakim Nivre, Yue Zhang, and Koldo Gojenola. 2014. On wordnet semantic classes and dependency parsing. In *Proceedings of the 52th Annual Meeting of the Association of Computational Linguistics*, pages 649–655, Baltimore (Maryland) - (USA), June. Association for Computational Linguistics.
- [7] H. Song, L. Chen, R. Gao, I. I. M. Bogdan, J. Yang, S. Wang, W. Dong, W. Quan, W. Dang, and X. Yu, "Automatic schizophrenic discrimination on fnirs by using complex brain network analysis and svm," *BMC Medical Informatics and Decision Making*, vol. 17, no. 3, p. 166, Dec 2017.
- [8] Iakes Goenaga, Koldo Gojenola, and Nerea Ezeiza. 2014. Combining clustering approaches for semisupervised parsing: the basque team system in the sprml2014 shared task. In *First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages: Shared Task on Statistical Parsing of Morphologically Rich Languages*, Dublin, Ireland, August. Dublin City University.
- [9] M. Jiang et al., "Text classification based on deep belief network and softmax regression," *Neural Comput. Appl.*, 2018.

- [10] G. Mujtaba, L. Shuib, R. G. Raj, R. Rajandram, and K. Shaikh, "Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study," *J. Forensic Leg. Med.*, 2018.
- [11] W. Hadi, Q. A. Al-Radaideh, and S. Alhawari, "Integrating associative rule-based classification with Naïve Bayes for text classification," *Appl. Soft Comput. J.*, 2018.
- [12] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Mark.*, 2019.
- [13] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. Muchaluat Saade, "A bayesian network decision model for supporting the diagnosis of dementia, alzheimer's disease and mild cognitive impairment," *Computers in Biology and Medicine*, vol. 51, pp. 140–158, 2017.
- [14] Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 809–815.
- [15] Kepa Bengoetxea, Eneko Agirre, Joakim Nivre, Yue Zhang, and Koldo Gojenola. 2014. On wordnet semantic classes and dependency parsing. pages 649– 655. *ACL 2014*.
- [16] Djame Seddah, Sandra K ¨ ubler, and Reut Tsarfaty. " 2014. Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland, August. Dublin City University.
- [17] J. K. Johannesen, J. Bi, R. Jiang, J. G. Kenney, and C. A. Chen, "Machine learning identification of eeg features predicting working memory performance in schizophrenia and healthy adults," *Neuropsychiatric Electrophysiology*, vol. 2, no. 1, p. 3, 2016.
- [18] P. Patel, P. Aggarwal, and A. Gupta, " Classification of schizophrenia versus normal subjects using deep learning," in *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, ser. *ICVGIP '16*. New York, NY, USA: ACM, 2016, pp. 28:1–28:6. [Online]. Available: <http://doi.acm.org/10.1145/3009977.3010050>

Turnitin Originality Report

Processed on: 20-Feb-2020 11:07 PM CST ID:

1261304755

Word Count: 4199

Submitted: 1

Similarity Index

16%

Similarity by Source

Internet Sources: 9%
Publications: 11%
Student Papers: 9%

A Unique & Dynamic
Approach to Predict
Schizophrenia Disease using

Machine Learning By
Neelisetti Ashok Tatikonda
Venkata Sai Manoj

4% match (publications)

[Subhan Tariq, Nadeem Akhtar, Humaira Afzal, Shahzad Khalid, Muhammad Rafiq Mufti, Shahid Hussain, Asad Habib, Ghufra Ahmad. "A Novel Co-Training-Based Approach for the Classification of Mental Illnesses Using Social Media Posts", IEEE Access, 2019](#)

2% match (publications)

[Bei Wei. "A case study of publishing atlas of remote sensing images", Proceedings 2005 IEEE International Geoscience and Remote Sensing Symposium 2005 IGARSS 05, 2005](#)

1% match (Internet from 25-Jan-2020)

<https://studylib.net/doc/25190944/daniel-jurafsky--james-h.-martin-speech-and-language-proc...>

1% match (student papers from 27-Jun-2019)

[Submitted to University of Western Macedonia on 2019-06-27](#)

1% match (student papers from 20-Feb-2020)

[Submitted to President University on 2020-02-20](#)

1% match (Internet from 04-Aug-2013)

<http://oru.diva-portal.org/smash/get/diva2:389393/FULLTEXT01.pdf>

1% match (Internet from 18-Dec-2019)

<https://denisakristianti.home.blog/2019/12/11/online-identity-2/>

< 1% match (Internet from 22-Oct-2014)

http://martinweisser.org/courses/intro/penn_tagset.html

< 1% match (student papers from 12-Nov-2017)

[Submitted to Higher Education Commission Pakistan on 2017-11-12](#)

< 1% match ()

<http://nrl.northumbria.ac.uk/38304/1/Li%20et%20al%20-%20A%20Novel%20Smart%20Energy%20Theft%20System%20%28SETS%20%20for%20IoT%20based%20Smart%20Home%20AAM.pdf>

< 1% match (publications)

[Signals and Communication Technology, 2015.](#)

< 1% match (student papers from 18-Aug-2015)

[Submitted to Universiti Kebangsaan Malaysia on 2015-08-18](#)

< 1% match (Internet from 22-Jan-2013) http://trac.sketchengine.co.uk/wiki/SkE/Config/FullDoc
< 1% match (student papers from 24-Apr-2019) Submitted to Aberdeen College on 2019-04-24
< 1% match (student papers from 23-Dec-2019) Submitted to Hofstra University on 2019-12-23
< 1% match (student papers from 24-Apr-2015) Submitted to Universiti Malaysia Sarawak on 2015-04-24
< 1% match (Internet from 12-Nov-2019) https://ijmce.com/resources/Paper-Format-IJMCE.doc
< 1% match (Internet from 29-Oct-2018) http://www.dtic.mil/dtic/tr/fulltext/u2/1056519.pdf
< 1% match () http://eksl.cs.umass.edu/~casutton/publications/lcrf.pdf
< 1% match (publications) Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, Khairunisa Shaikh. "Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study", Journal of Forensic and Legal Medicine, 2018
< 1% match (Internet from 02-Sep-2019) https://www.mdpi.com/1996-1073/12/17/3310
< 1% match (publications) "Advances in Computational Intelligence", Springer Science and Business Media LLC, 2019
< 1% match (publications) Deepak Agnihotri, Kesari Verma, Priyanka Tripathi. "Variable Global Feature Selection Scheme for automatic classification of text documents", Expert Systems with Applications, 2017
< 1% match (Internet from 02-Apr-2009) http://www.se.cuhk.edu.hk/~seg5470/lecture/POS-HMM.pdf
< 1% match (publications) Hector Allende-Cid, Juan Zamora, Pedro Alfaro-Faccio, Maria Francisca Alonso-Sanchez. "A Machine Learning Approach for the Automatic Classification of Schizophrenic Discourse", IEEE Access, 2019
< 1% match (student papers from 04-Sep-2017) Submitted to The University of Manchester on 2017-09-04
< 1% match (Internet from 14-Aug-2018) http://ijarcet.org/wp-content/uploads/IJARCET-VOL-5-ISSUE-1-30-34.pdf
A Unique & Dynamic Approach to Predict Schizophrenia Disease using Machine Learning Detection of Schizophrenia Discourse from Twitter Tweets 1st Neelisetti Ashok CSE Department Sathyabama University Chennai, India neelisettiashok@gmail.com Tatikonda Venkata Sai Manoj

CSE Department Sathyabama University Chennai, India

saimanoj1010@gmail.com Abstract— The society environment and its

wide array of influences factors are very much connected with each

phase of human life operating and development. Schizophrenia is a

chronic and severe mental disease which alters the human memory and mind.

Mental fitness and mental disease are dictated by various and collaborating

social, mental, what's more, organic variables, similarly as wellbeing and

sickness when all is said in done. The clearest proof identifies with the

dangers of mental ailments, which in the created and creating world are

related with markers of neediness, including low degrees of training. The

singular goal of this proposed project is to automatically distinguish patients

with schizophrenia and without schizophrenia. Detection of schizophrenia in

initial periods is crucial to prevent comprehensive mental illness. Currently,

identifying the schizophrenia discourse which are extensively distributed in

our culture and forecasting the forthcoming phases of the mental illness has

become extremely crucial in modern-day society. Normally patients

express their health condition in [one of the most popular social media](#)

networks called [Twitter](#). We introduce a unique and dynamic machine

learning (ML) approach to predict the schizophrenia discourse from

twitter tweets. This unique approach is extremely adaptable as

practically any supervised machine learning algorithm can be utilized for

prediction. The efficiency and accuracy of the machine learning

algorithm is directly proportional to the excellence of the feature

engineering which needs a high level of domain knowledge. To

overcome this problem, this paper proposes [a Deep Neural Network](#)

[\(DNN\)](#) which [is](#) capable [of](#) automatically identifies [the](#) pattern of

schizophrenia discourse and classifies the schizophrenia discourse.

Keywords — Machine Learning, Deep [Learning](#), [Convolutional Neural](#)

[Network CNN](#), [Long-Short Term Memory LSTM](#), [Natural Language](#)

[Processing](#), Discourse Comprehension, [Schizophrenia I. INTRODUCTION](#)

[Schizophrenia is a](#) dangerous neurological sickness, which seems in

roughly 1% of the adult population around the globe. It affects

remembrance, awareness, linguistic and executive functions of the

brain. According to World Health Organization (WHO) stated that the

worldwide incidence of global Schizophrenia is just about 20 million

people. Among that 69% percentage of Schizophrenia disorder patients

are not obtaining proper medications. There are various causes for

Schizophrenia disorders. Any research does not distinguish one

particular aspect. Hereditary and environmental influences perform a

major role in its diagnosis. It is normally developed in the adulthood

under the age of 26. The patients which are affected by this disease will

have intellectual deficiencies, including reading disorders. They will have

difficulty to select suitable phrases while they are talking and

composing. It modifies patients' observation, points of view, and

conduct as confirmed by fantasies, dreams, muddled discourse or conduct,

social withdrawal, and differed intellectual deficiencies. Fig 1.

Schizophrenia Word Cloud Schizophrenia Discourse (SD) is a mental

ailment, which is the highly predominant neurodegenerative disease in

this globe. It is a radical and dense mental illness influencing

susceptible peoples, whose pervasiveness is projected to twice by 2030.

Diagnosis and medication of neurological Discourse are emphases in

brain study. Despite the fact that meds have been created to treat

schizophrenia, they don't work for all patients, and there is proof that a

few types of schizophrenia seem, by all accounts, to be treatment safe.

This just underlines the significance of leading fruitful schizophrenia

clinical preliminaries. According to the artificial intelligence study, social

media text analysis has been a leading area of study in the past two to

three decades that has been commonly applied in wide array [XXX-](#)

X-XXXX-XXXX-X/XX/\$XX.00 ©20XX IEEE of research such as healthcare, smart home, surveillance systems, human-computer interaction, gaming and so on. With an exceptional development of web-based social networking, huge number of individuals deliberately share a lot of information by communicating their states of mind, sentiments, feelings, and day by day battles with emotional well-being issues via web-based networking media stages like Twitter. This offers open doors for new comprehension of these networks. With the improvement of Internet utilization, individuals have begun to impart their encounters and difficulties to emotional wellness issue through online discussions, small scale web journals or tweets. Their online exercises roused numerous specialists to present new types of potential human services arrangements and techniques for early melancholy recognition frameworks. Utilizing distinctive Natural Language Processing (NLP) methods and content characterization draws near, they attempted to prevail in a better improvement. The singular aim of this application is to develop a machine learning especially deep learning approach, named Long-Short Term Memory (LSTM) plus Convolutional Neural Network (CNN), for detecting Schizophrenia disorders using features obtained from a different dataset from kaggle.com, namely the Schizophrenia Discourse dataset. The proposed LSTM+CNN is competent to handle high-ranking dimensional data by expanding the model interpretability and discovering the highly appropriate features while dispensing with the typical inequitable class division. Natural language processing (NLP) has gained significantly from the revival of neural networks (NNs), expected to their superior execution with fewer necessary of contrived features. Convolutional neural network (CNN) and recurrent neural network (RNN) are the two major types Deep neural network architecture. Gating systems have been established to improve certain constraints of the basic recurrent neural network, causing in two fundamental recurrent neural network types: long short-term memory (LSTM) and gated recurrent unit (GRU). This application is therefore dedicated on approaches to categorize symptoms of Schizophrenia Discourse in posted tweets as early as feasible established on deep learning methodology. This paper content is organized into six divisions. The first division contains the Introduction section. The second division specifies about social network. The third division illuminates related work. Our proposed system is deliberated in the fourth division which covers the datasets, data analysis procedure, the proposed LSTM+CNN approach. The experiment results are presented in the fifth division and conclusions are illustrated in the final segment. The final segment also contains the discussion as well as the future work in that direction.

II. SOCIAL NETWORK

The social media network has become immeasurable and reputation along with all the service industries today. Today most of the people around the world are using the social media network to share their individual, qualified interest and condition consciousness. Social media networks are the collection of collaborative web-based technology where interested communities produce and share expertise, concepts, intelligence, specialized interests and other information through virtual communities using online communication channels. There are various types of social medias like wikis, Social Networks, Media Sharing Networks, Discussion Forums, Bookmarking & Content Curation Networks, Social Review Websites, Community Blogging & Publishing Websites, Image & Video Hosting Networks, Sharing Economy Networks, Social Shopping Networks, Interest-Based Networks, Anonymous Social Networks and many more. The popular social networks are Facebook, Twitter, LinkedIn etc. The social networks are web based social media applications which are constantly evolving. The communities uses the

social network to connect people and share knowledge through human-to-human interaction. It is also called as relationship networks where like-minded people can share ideas and knowledge. The popular image and video hosting websites are Flickr, photobucket, Instagram, Snapchat, YouTube, Vimeo etc. It is also called as Media Sharing Websites. These media sharing networks can be used to share photos, images, videos, live videos, audios and other medias which can be accessed from anywhere and anytime around the world. The image and video hosting websites can be used to promote brand recognition, lead creation etc. These networks mostly promote user generated content. The widespread Discussion Forums networks are reddit, Quora, Digg, Warrior Forum, etc. Discussion forum are intrinsically social; since it brings community people collectively on a collective ground. In these networks, people participate in discussions by publishing and replying to community messages and also share news, information and other options. The discussion forum can be either public or private depending on the conversation intentions. These networks are highly effective for the market research team. The predominant social review networks are TripAdvisor, Yelp, FourSquare, Google My Business, BBB (Better Business Bureau) etc. These networks allow the online visitors to determine, share and review for all sorts of location and experiences like services, products and brands as well as restaurants, travel destinations, and just about anything else. These networks are highly effective to understand the customer's and consumer's perspectives. It also provides a valuable feedback for any kind of businesses. The common Community Blogging & Publishing Websites are Medium, Tumblr, Blogger, Wordpress etc. A blogging platform is a web-based application which can be used to publish your creative content online. These network can be used to publish and find the online content. It is the most successful methodology to engage with your customers or audience. Few Blogging networks are exclusive. This means that it can be only accessed by invite. The widespread Sharing Economy Networks are Airbnb, Pantheon, Kickstarter etc. These kind of networks can be used to promote, discover, distribute, purchase, trade and export & import products and services between community. It is also called as peer economy or gig economy network. Few online marketplaces offer tourism experiences. The predominant Bookmarking & Content Curation Networks are Pinterest, Flipboard, Digg, StumbleUpon etc. The online visitors can use these Bookmarking & Content Curation Networks to save and organize hyperlinks to any number of web resources and web pages. These applications allow the users to categorize the web links which can be shared among their followers. These categories and web links can be easily searchable. These networks can be used to discuss new and trending messages, images and videos. The primary focus on Content Curation Networks are to determine and share trending articles and other text content.

III. RELATED WORK Most of the existing Natural Language Processing framework such as Stanford CoreNLP and SpaCy are only concentrates on predicting semantic forms rather than modeling Natural Language Processing (NLP) designs. Many of the recent research studies have explored linguistic expressions of Schizophrenia Disease in social network dataset especially from Twitter. Some of the recent research studies focus on mental illness and their moods and some of the other research papers focus on Schizophrenia Disease classifications. Attention Discrepancy Hyperactivity Syndrome [2] classifies the patients correlated with chronic Schizophrenia Disease. In this research, the dataset is extracted from the Reddit which is a famous social network community platform. This paper uses the Reddit API to develop the corpus. It implemented using the co-training methodology with the widely used classifier algorithms like Support

Vector Machine, and Naïve Bayes. In the study [4] detects the mental health expression and where the corpus is extracted from college student influenced data from Reddit API. This paper analysis the verbal and chronological attributes of the Schizophrenia Diseases from college students in online social media communities. It implements a novel inductive transfer learning approach. It applies the binary machine learning classification algorithm with positive and negative as class labels. The machine learning classification algorithm like logistic regression, random forest, Ada Boost etc are used for classifying the Schizophrenia Diseases. In research study [9], inventors used Deep Beliefs Networks (DBN) which adopts a layer by layer training approach. The dataset contains high thin dimensionality feature vectors. Deep Beliefs Networks (DBN) is used solving this high dimensionality problem and also decreases the computational complexity from extracting the high level feature to low level features. The classification of the data is implemented using multinomial logistic regression. multinomial logistic regression is a multi-class classification framework. The outcome of this study is better than Support Vector Machine and Neural Network. In multinomial logistic regression, the sigmoid logistic function is replaced by SoftMax function. This research defines the net input $z = w_0x_0 + w_1x_1 + \dots + w_mx_m = \sum_{i=0}^m w_i x_i = w^T x + b$. In this research [10], authors perform the automatic text classification from the medical documents. Dataset are gathered from medical hospital which contains the information about 8 different reasons of death. In the preprocessing steps, authors change the corpus to the lower case statements. They have also removed the special characters, punctuations and stop words using the Natural Language Processing Toolkit (NLTK). After that, unigram, bigram, trigram, and hybrid-gram features were obtained from the pre-processed corpus. In this manner, creators have utilized distinctive best-in-class content orders models, for example, SVM, NN, NB, and Random Forest (RF). At long last, creators detailed the SVM as an outflanked classifier when contrasted with others. In study [12], creators have performed a research to evaluate the accomplishment of wide range of techniques for automatic conceptual overview of text classification. We recognize opinion and substance grouping, which are both quite compelling for promoting applications. In short order, the objective of the content grouping strategies is distinguish the feeling passed on through this content and to effectively order it as negative. Dictionary based or regulated AI strategies are the two significant ways to deal with achieve this errand. The author has used five machine learning algorithms 1. Artificial Neural Network (ANN) 2. K-Nearest Neighbours (KNN) 3. Naïve Bayes (NB) 4. Random Forest (RF) and 5. Support Vector Machine (SVM). The authors used five lexicon based methods such as 1. AFINN-en-165.txt 2. Bing Liu and collaborators 3. nrc from Saif Mohammad and Peter Turney 4. ADER (Valence Aware Dictionary for sEntiment Reasoning). The first three of these lexicons are based on unigrams, i.e., single words. Given the above investigations, we can reason that SVM, NB and Random Forest perform better for classification of content. It fundamentally relies upon the attributes of a Dataset which is under investigation. The vast majority of the investigations utilized basic SVM or SVM with some alteration so as to show signs of improvement classification accuracy. Likewise, for RF it tends to be deduced that it is generally excellent for multi-class classification and NB additionally functions admirably with the dataset of little example sizes. Subsequently, the Machine learning approaches beat the dictionary based calculations. Besides, it tends to be seen that feature extraction and feature determination procedures can help improve the classification accuracy. IV. PROPOSED WORK After the text editing has been completed, the paper is ready for the template. Duplicate

the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Data

Acquisition To train the machine learning algorithm, we require two different types of datasets: 1. Schizophrenia discourse dataset 2. Non-Schizophrenia discourse dataset. One of the major steps for neuroscience is Data gathering or Data Acquisition; since a minor error in the data acquisition may possibly trigger imprecision in prediction analysis after the predictive model is built. Twitter is a web based micro blogging and social networking service provider in which online visitors can post and obtain messages which can be the maximum size of 140 characters. In Twitter, these messages are referred as tweets. The Kaggle.com already collected the tweets related to Schizophrenia discourse and provides us to download those tweets as datasets. To acquire these clients, Kaggle.com arbitrarily inspected Twitter IDs. From that point, we continued to download the tweets from the chosen IDs. Sample Schizophrenia Dataset Sample Non-Schizophrenia Dataset Tweets Summary B. Preprocessing This research only contemplates the English authored subject. The next phase of the proposed approach is

data preprocessing. In any machine learning algorithms, the preprocessing of the dataset normally contains various steps. Step 1: Rephrase all tweeter seeds as phrases or unigrams. Step 2: Delete all retweet handlers Step 3: Delete all the URL from each tweets Step 4: Delete all Non-ASCII characters and words from the Schizophrenia discourse dataset and Non-Schizophrenia discourse dataset Step 5: Break Usernames and Hashtags Step 6: Delete any punctuations from the hashtags. Step 7: Delete all punctuations from the tweets Step 8: Load the prebuilt datasets and remove those stop words from datasets Step 9: Delete all leading and trailing spaces. Step 10: Delete all numbers from the tweets Step 11: Delete all multiple in between text white spaces from the tweets C. Natural language Processing (NLP)

Natural language processing provides efficient methodology to investigate the text content. Inheritance Structures and Significance of the text content can be determined using the Syntax analysis and Semantic analysis. One of the major tasks for NLP is to classify the sequence. In other words, it is used to designate the categorical label for each word in the text content. This can be implemented using 1)

Part of Speech Tagging 2) Named Entity Recognition Part of Speech Tagging Parts of Speech (POS) Tagging can be used for information retrieval. There are two main kinds of tagsets available in Natural

Language Processing 1. Brown Corpus Tagset 2. Penn Treebank Tagset Few of the Brown Corpus Tagset Tag Description CC Coordinating Conjunction CD Cardinal Numeral CS Subordinating Conjunction DT singular determiner DTI singular or plural determiner/ quantifier DTS plural determiner DTX determiner/double conjunction FW foreign word HL word occurring in headline IN Preposition JJ Adjective JJR Comparative Adjective JJS Semantically Superlative Adjective JJT Morphologically Superlative Adjective NN Singular Or Mass Noun Few of the

Penn Treebank tagset Tag Description CC Coordinating Conjunction CD Cardinal Numeral EX Existential There IN Preposition, Subordinating Conjunction POS Possessive Ending PP Personal Pronoun PP\$ Possessive Pronoun RB Adverb RBR Adverb, Comparative RBS Adverb, Superlative SENT Sentence-Break Punctuation VB Verb Be, Base Form POS Labeling is one of the apparatuses and the segments for foundation to Natural

Language Processing of a given language. POS labeling is vital in numerous fields, for example, content expression, language structure, semantic investigation and interpretation A tagset is a vector of parts of

speech tags. It designates every single word the extremely expected within the framework. There are three methodologies are available for Parts of Speech (POS) Tagging. 1) Rule-Based POS [tagging](#) 2) [Transformation-based tagging](#) 3) [Stochastic \(Probabilistic\) tagging](#). Let us consider $W = w_1 w_2 w_3 \dots w_n$ Series of words $T = t_1 t_2 t_3 \dots t_n$ Series of POS Tags $f: W \rightarrow T = f(W)$ Rule based POS Tagging Calculation for allocating grammatical feature dependent D. The Proposed Architecture on two phases, the main stage utilizes a vocabulary (lexicon) to dole out each word a rundown of potential grammatical features, the subsequent stage utilizes an enormous rundown of manually [writtendisambiguationrules to winnowdownthis](#) [rundown](#) [to a](#) solitary grammatical feature for each word Data Sets Chennai is a nice place . But weather is so bad in Chennai [Adjective JJ](#) [Adverb RB](#) [Conjunction CC](#) [Determiner DT](#) [Noun NN](#), [NNP](#) [Preposition IN](#) [Verb VB](#) [ZTextTokenization](#) [Determine Phrases in Vector Space](#) [Part of Speech Tagging](#) [Named Entity Recognition](#) [Named Entity Recognition](#) (NER) can perform [the](#) process of entity identification and information retrieval. It is capable of classifying the named entities which are present in the text content into pre-defined categories like people, locations, organizations, percentages, quantities, countries, dates, individuals, product information, service information etc. It provides a solid semantic meaning for the text content so that it is easily to understand the subject of any text content. Named Entity Recognition (NER) is a sub issue of data extraction and includes handling organized and unstructured records and distinguishing articulations that allude to people groups, spots, associations and organizations. NER is an essential errand and it is the center of regular language handling (NLP) framework. NER includes two assignments, which are: right off the bat the recognizable proof of appropriate names in content, and also the classification of these names into a lot of predefined classes of intrigue, for example, persons, organizations, areas, date and time articulations

Shallow Parsing Semantic Analysis Classification V. EXPERIMENT RESULTS Performance Evaluation Class Precision Performance Measures [Recall F-measure](#) Anxiety [0.89](#) [0.80](#) [0.82](#) ADHD [0.66](#) [0.67](#) [0.62](#) Depression [0.84](#) [0.87](#) [0.81](#) Bipolar [0.60](#) [0.80](#) [0.70](#) From the above table, we can monitor the efficiency of suggested methodology [in terms of class wise](#) categorization [of](#) tweeter statements. [In terms of F-measures, we](#) monitor [the](#) accomplishment [of](#) recommended method [for the](#) categorization [of Anxiety](#) tweets [and its related comments \(F-measure = 0.82\)](#) [as compared to ADHD \(F-measure = 0.62\), Depression \(F-measure = 0.84\), and Bipolar \(F-measure = 0.70\).](#)

Summarize history for accuracy Summarize history for loss VI.

CONCLUSION & FUTURE WORK In this novel system, we implemented an artificial intelligence application to division normal and abnormal Schizophrenia subjects from social media networks using machine learning methodology. This work proposes a LSTM+ CNN based framework for automated Schizophrenia Discourse detection, to possibly gather connection phrases from tweets text contents. The semantic consistence between the machine and human affiliation systems is tentatively confirmed. This discovering gives a knowledge of comprehension the affiliated. From the all statements described in this paper, we can easily conclude that all the phases engaged in this approach are important and they all impact to obtaining the finest out of the algorithms. This complete procedure stream developed in additional precise analysis prediction of Schizophrenia discourse. In the future, we would like to extend this project by analysing for other diseases in the same mental illness areas.

REFERENCES [1] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. Muchaluat Saade, -A bayesian network decision model for supporting the diagnosis of dementia, alzheimers

disease and mild cognitive impairment, *Computers in Biology and Medicine*, vol. 51, pp. 140–158, 2017. [2] Subhan Tariq, Nadeem Akhtar, Humaira Afzal, Shahzad Khalid, Muhammad Rafiq Mufti, Shahid Hussain, Asad Habib, Ghufraan Ahmad. A novel Co-training based approach for the classification of mental illnesses using Social media posts 2019 [3] Rongjun Yu, Yi-Ling Chien, Hsiao-Lan Sharon Wang, Chih-Min Liu, Chen-Chung Liu, Tzung-Jeng Hwang, H. Ming Hsieh, Hai-Gwo Hwu, and Wen-Yih Isaac Tseng, -Frequency-Specific Alternations in the Amplitude of Low-Frequency Fluctuations in Schizophrenia, *Human Brain Mapping*, vol. 35, 2014, pp. 627–637. [4] S. Bagroy, P. Kumaraguru, and M. De Choudhury, -A Social Media Based Index of Mental Well-Being in College Campuses, pp. 1634–1646, 2017. [5] E. Allen, E. Damaraju, S.M. Plis, E. Erhardt, T. Eichele, V.D. Calhoun, -Tracking whole-brain connectivity dynamics in the resting state, *Cereb. Cortex*, vol. 24(3), 2013, pp. 663–676, 2013. [6] Eneko Agirre, Kepa Bengoetxea, Joakim Nivre, Yue Zhang, and Koldo Gojenola. 2014. On wordnet semantic classes and dependency parsing. In *Proceedings of the 52th Annual Meeting of the Association of Computational Linguistics*, pages 649–655, Baltimore (Maryland) - (USA), June. Association for Computational Linguistics. [7] H. Song, L. Chen, R. Gao, I. I. M. Bogdan, J. Yang, S. Wang, W. Dong, W. Quan, W. Dang, and X. Yu, -Automatic schizophrenic discrimination on fNIRS by using complex brain network analysis and svm, *BMC Medical Informatics and Decision Making*, vol. 17, no. 3, p. 166, Dec 2017. [8] lakes Goenaga, Koldo Gojenola, and Nerea Ezeiza. 2014. Combining clustering approaches for semisupervised parsing: the basque team system in the spmrl2014 shared task. In *First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages: Shared Task on Statistical Parsing of Morphologically Rich Languages*, Dublin, Ireland, August. Dublin City University. [9] M. Jiang et al., -Text classification based on deep belief network and softmax regression, *Neural Comput. Appl.*, 2018. [10] G. Mujtaba, L. Shuib, R. G. Raj, R. Rajandram, and K. Shaikh, [16] Djame Seddah, Sandra K. Kubler, and Reut Tsarfaty. "2014. -Prediction of cause of death from forensic autopsy reports using text Introducing the spmrl 2014 shared task on parsing morphologically- classification techniques: A comparative study, *J. Forensic Leg. Med.*, rich languages. In *Proceedings of the First Joint Workshop on 2018. Statistical Parsing of Morphologically Rich Languages and Syntactic* [11] W. Hadi, Q. A. Al-Radaideh, and S. Alhawari, -Integrating associative Analysis of Non-Canonical Languages, pages 103–109, Dublin, rule-based classification with Naïve Bayes for text classification, Ireland, August. Dublin City University. *Appl. Soft Comput. J.*, 2018. [17] J. K. Johannesen, J. Bi, R. Jiang, J. G. Kenney, and C. A. Chen, [12] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, -Comparing-Machine learning identification of eeg features predicting working automated text classification methods, *Int. J. Res. Mark.*, 2019. memory performance in schizophrenia and healthy adults, [13] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. Muchaluat Neuropsychiatric Electrophysiology, vol. 2, no. 1, p. 3, 2016. Saade, -A bayesian network decision model for supporting the [18] P. Patel, P. Aggarwal, and A. Gupta, -Classification of schizophrenia diagnosis of dementia, alzheimers disease and mild cognitive versus normal subjects using deep learning, in *Proceedings of the impairment, Computers in Biology and Medicine*, vol. 51, pp. 140–Tenth Indian Conference on Computer Vision, Graphics and Image 158, 2017. Processing, ser. ICVGIP '16. New York, NY, USA: ACM, 2016, pp. [14] Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring 28:1–28:6. [Online]. Available: continuous word representations for dependency parsing. In <http://doi.acm.org/10.1145/3009977.3010050>

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22 -27, 2014, Baltimore, . MD, USA, Volume 2: Short Papers, pages 809 –815. [15] Kepa Bengoetxea, Eneko Agirre, Joakim Nivre, Yue Zhang, and Koldo Gojenola. 2014. Onwordnetsemantic classes anddependency parsing. pages 649– 655. ACL 2014.