Analysis and Derivation of Optimum Domain Specific Semantic Model for Detecting Depression Text From Twitter Stream

by Bhaskar Bhakat

Submission date: 01-Jun-2021 05:06AM (UTC+0900)
Submission ID: 1597788052
File name: G_16_Final_report_for_check.docx (1.14M)
Word count: 7984
Character count: 43155

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Sentiments are depicted as genuine slants that are focused on someone or something considering inside or outside events having explicit significance for the individual. Also, the web, today, has got a vital medium through which individuals express their feelings, sentiments, and assessments. Each occasion, news, or movement all throughout the planet, is shared, examined, posted, and remarked via web-based media, by a great many individuals.

Sadness causes negative thinking, less concentration in work, diminished proficiency. It furthermore impacts the proliferation course of action of the individual. Gloom may create mental turmoil besides. Early acknowledgment of emotional well-being is handily treated at a beginning phase. Mental prosperity handles pressure and is an activity for settling on decisions in each step for the duration of regular daily existence.

Emotional well-being plays a significant factor in each phase of life whether it be pre-adulthood or an adult. Mental prosperity of the pressing factor at home and workplace. It will fabricate the benefit of people.

In this paper, we propose a strategy to arrange and measure tweets as per six standard feelings that are well-known to everybody. Here, we base our examination on tweets posted on Twitter, yet it tends to be handily stretched out to any sort of text whether it is one-lined features, messages, and posts via online media or bigger lumps of compositions, due to programmed improvement of our preparation set.

While the exactness is higher than the review for the best tweet models and the review is higher than the accuracy for the best content models, the best models for the two modalities have equivalent exactness. Regardless of better execution, the best content models require less highlights than the best tweet models based on highlights from a similar transient amount of information (except for 56 days of information).

This paper depicts a bunch of examinations performed on open tweets, to distinguish clients who experience the ill effects of gloom or are in danger of discouragement, utilizing text mining strategies. We additionally endeavor to distinguish troubled tweets in a Twitter stream.

We have developed a structure that could score and name any piece of text, especially tweets and posts by means of online media as appeared by six Emotion-Categories: Happiness, Sadness, Fear, Surprise, Anger and Disgust close by their force scores, utilizing its scholarly features, a grouping of NLP gadgets and standard Machine Learning classifiers. Another colossal responsibility is that we have successfully arranged a system that could normally (with no manual effort) collect a gainful planning set for our ML Classifiers, involving a gigantic enough course of action of stamped tweets from all Emotion-Categories. We have made a tremendous bunch of words in English, that contains words conveying a particular inclination close by the force of that inclination. We have achieved an accuracy of about 94%.

1.2 OBJECTIVE

The objective of this hypothesis is to utilize Natural Language Processin g and Machine Learning philosophies to fabricate a structure that can recognize in peril tweets and, subsequently, in peril customers given a bunch of tweets from a customer. We need to distinguish intriguing abstract or neighborhood for this task. This structure should represent data got through online media posts, for example, 1) tweets are short and may pass on less feeling; 2) customers tweet about a wide scope of subjects; and 3) the language doesn't adjust to syntactic plan and may contain spelling botches/shorthand documentation to find a way into 140 characters.

This prompts an auxiliary objective of the proposal, which is to recognize important tweets for investigation from the huge measure of Twitter information.

1.3 PROBLEMSTATEMENT

To analyze the emotions and sentiments of users or for a particular domain from twitter data. Also, to detect depression of a particular user using real time user twitter dataset.

1.4 ORGANIZATION OF REPORT

The entire construction of undertaking has been cut up in consecutive methodology. There will be itemized outline of the task with the utilization of constant informational indexes including the information of patients. Exploratory twisting investigation of information is altogether concentrated in recognizing conduct of people over the long haul on.

Furthermore, we are proposing various devices required for performing the prosperity techniques linked to dataset. Healing assumptions is the major segment when perspective for exploring the sum covering access for result of our model.The surviving from the outcome is talked about below:

1.Literature Survey and Method of the expectation model.
2.Calculation and scientific methodology
3.Algorithm execution
4.Results, Dataset in Testing planning for the task.
5.Results exactness investigation, calculation of precision in information expectation alongside stable structure climate for the undertaking.

CHAPTER 2
LITERATURE SURVEY

2.1 INTRODUCTION

Sorrow and other mental illnesses are hard to distinguish, bringing about a high pace of underdiagnosis. This is expected to some extent to the absence of a lab. Most mental infections can be tried for. Psychological maladjustment is hard to analyze. in light of self-revealed encounters, family members' reports, and an overview Examine your psychological state. This section centers around three principle regions where we led research. To start with, we survey the requirement for a framework that can help in the location of mental infections. Second, we evaluate the legitimacy of web-based media as an information source. At last, we take a gander at momentum strategies that have been created to resolve comparable issues, for example, diagnosing mental illnesses utilizing online media.

The need for a framework that can help experts in recognizing in danger people on a major scale was set up in this part. We at that point found that online media is a generally utilized and dependable wellspring of information for study and that it could be used to analyze psychological instability. We at that point checked that text mining can be utilized to analyze dysfunctional behavior via online media with respectable precision.

We will use Twitter information to check whether those in danger of misery can be distinguished utilizing these significant components in our examination. We'll see whether stressed tweets can be spotted utilizing text-mining instruments later on. We should call attention to that deciding if a psychological illness requires treatment is a troublesome expert judgment. It's additionally important that deciding the seriousness of the condition is an intense cycle that must be refined by a profoundly prepared master utilizing an assortment of approaches like composed depictions, clinical meetings, and their own judgment (APA, 2013).Detecting in danger clients to psychological maladjustment inside web-based media utilizing web mining and feeling investigation strategies could be viewed as an essential advance in making an engaged gathering of populace, given the more noteworthy intricacy of the techniques and level of abilities associated with recognizing mental confusion and the fundamental therapies. Recognizing mental problems via online media won't be a stage for naming a person as a

patient with a particular issue, but instead a stage for raising an alert so the proper specialists can do whatever it takes to additionally research the anticipated client to affirm their psychological wellness status.

2.2 EXISTING STATEMENT

Depression is the most common cause of disability, yet it is also one of the most curable mood disorders. As a result, unobtrusively detecting depression is critical. Many studies are beginning to use machine learning for depression detection based on social media and smartphone data to replace the conventional survey tools used to screen for depression. We assess the ability of a privately available vs a publically available method to screen for depression in this study.

We use text messages and tweets from the past two weeks to a year to predict scores on the Patient Health Questionnaire-9, a popular depression screening tool. This is the first research to use the combined Moodable and EMU datasets' retroactively gathered crowd-sourced texts and tweets. Comprehensive feature engineering, feature selection, and machine learning are all part of the current methodology.

Word category frequencies, part of speech tag frequencies, emotion, and loudness are among the current 245 characteristics. Logistic Regression, which is based on the top 10 characteristics from two weeks of text data, is the best model. The average F1 score for this model is 0:806, the AUC is 0:832, and the recall is 0:925.

2.3 LITERATURESURVEY

In The Literature survey examines the different models completed by different creators in such a way that aides in the identification of Depression. The strategy uses various strategies identified with AI computations unmistakably and achieves high effectiveness in their specific fields. The total of the writing audits that we have gone through is examined in a word underneath:

Guozheng Rao [1] 2020 proposed a model in which they were identifying Depressed Individuals in Online Forums. Their key aim was to perceive despairing

using the best immense neural arranging from two of the most standard basic learning approaches in normal language dealing with(CNNs) and (RNNs).

Mark E. Larsen [2] in 2015 conducted a study named as Mapping Emotion on Twitter. His study rehearses and astoundingly mixes clear audit relations with semantic and social signs amassed from Google Play application information (87K applications, 2.9M examinations, and 2.4M intelligent people, assembled over a goliath piece of a year), to see sketchy applications. It accomplishes over 95% exactness referring to the best quality level datasets of malware, deceiving, and genuine applications.

Marcel Trotzek [3] in 2018 conducted a study mainly known for Detection of Depression Indications in Text Sequences that evaluate using AI computations. These are customary in regular day-to-day existence where choices are taken by interlinked various rules. In the proposed structure, AI masterminds a tendency from different viewpoints and evaluation words. For example, in a bistro audit, the master cherishes the food yet disdains the help. The computation depicts the audit by thought words or verbalizations about viewpoints.

BudhadityaSaha, Thin Nguyen, Dinh Phung, and Svetha Venkatesh [4] in 2016 conducted a study that shows 75% of the conspicuous malware applications share searching for rank mutilation. Fairplay finds diverse lie applications that advantage right currently keep away from Google Bouncer's receptiveness improvement. Fairplay also helped the openness of more than 1,000 outlines, verifiable for 193 applications, that uncover another sort of "coercive" design crusade: clients are meddling with into framing positive assessments and present and study various applications.

Michael M. Tadesse [5] in 2019 made a comparative study. In this paper, they have done a near examination among five methodologies say TF-IDF, Naive bayes, LSTM, Logistic Regression, Linear help vector. Among every one of the 5 strategies we have tracked down that Long Short Term Memory(LSTM)- RNN has the most noteworthy exactness to recognize the burdensome tweets from twitter.

ML Tlachac [6] tried to demonstrate that reflectively collected content messages have an incredible potential when evaluating for wretchedness, more so than freely posted tweets. Utilizing only the earlier fourteen days of instant messages, our strategic relapse models foresee a twofold PHQ-9 score at the moderate gloom cutoff of 15 with a normal F1 score of 0.806 and AUC of 0.832.

Les Servi and Sara Beth Elson [7] consolidating a robotized text examination program with another numerical way to deal with distinguish shifts in Twitter clients' passionate articulations, the work introduced in this paper offers a chance to utilize the information got from web-based media to acquire equitably determined experiences into the elements of Twitter clients.

FirojFattulalShahare [8] in 2017 concluded an expansive viewpoint on organizing compulsion and propose an orchestrating bowing perceiving check structure for versatile Apps. In particular, we from the start proposal to absolutely find the planning craftiness by mining the surprising time frames, to be express driving get-togethers, of flexible Apps. We can use such driving social affairs for seeing the nearby inconsistency instead of a general attribute of App rankings. [9] They have research three sorts of requests, i.e., coordinating based certifications, rating based confirmations, and study based authentications, by showing Apps' engineering, rating, and study practices through quantifiable hypotheses tests. Proposed an advancement gathering approach to work with everyone in assertion for guile attestation. At last, they assess the proposed structure with veritable App information accumulated from the iOS App Store for quite a while period. In starters, we embrace the savvy instinct of the proposed plan and show the versatility of the disclosure computation furthermore as some retinene of planning cheating works out.

Norah Saleh Alghamdi [10] inspected Depression Symptoms in an Arabic Psychological Forum through Machine Learning model which uses Naïve Bayes algorithm. The examination shows an precision of 89%; recall 0.99; F1-score 0.946; AUC 90%.

Table 2.1 Existing models to Detect Depression from social media

Reference	Model	Features	Results
Choudhury, Gamon, Counts, Horvitz (2013)	SVM	engagement, ego-network, emotion, linguist. Style, dep. Language, demographics	accuracy: 70%; precision: 0.74
Tsugawa et al (2015)	SVM	features obtained from user activities can be used to predict depression; topics of tweets estimated with topic model are useful features; approximately 2 months of observation data are necessary for recognizing depression, longer observation periods may worsen accuracy.	accuracy: 69%
Schwartz et al (2014)	regression	Identifies the rate of change of depression over seasons. n-grams, linguistic behavior and LDA topics.	
Yu and Ho (2014)	SVM	Bag of Words, LSA, ICA to obtain multiple emotion labels	accuracy: 65%
CLPsych 2015			
Resnik et al. (2015)	SVM	supervised LDA, supervised anchor, lexical TF-IDF and a combination of all	precision above 0.80
Preotiuc-Pietro et al (2015b)	LinSVM	Bag of words, topics derived from clustering methods, meta data from user's profile	precision: 0.867

Respecting ethical aspects of the usage of social media data and its privacy is a major problem. We take precautions not to share the acquired data in accordance with Twitter's policy. We also ignore any identifiable user information for this task. To minimise additional hardship, we protect users' privacy and ethical rights.

CHAPTER 3

SYSTEM REQUIREMENT

3.1 INTRODUCTION

Unmistakably characterized prerequisites are get defined by the basic results which prompt tasks to an efficient manner. It sets in an adequate arrangement that is required between a consumer and a manufacturer when both of them are trying to achieve the same goal. Superior grade, count of requirements likewise alleviate monetary risks which keeps ventures for a given time frame.The business analytics body defines knowledge as requirements needed to fulfill our need.

The initial requirements is a difficult taks to implement since a bunch of cycles like eliciting, examining, detail, approving and boarding. The report examines principle sorts in prerequisites in coding and designing and helps us find better solution for getting themutilized.

3.2 SOFTWARE SPECIFICATION

3.2.1 PYTHON

The py project provides with a unlicensed and no cost scripting and coding language. Accordingly, we shall easily define python here where we can start working developing over the language. You could even give our own code for the contribution in the local area and python development community. Along with this,this language is also a cross platform viable programming language. Anyway, what's the significance here? Apart from all this we can even execute or run python codes via open sourced frameworks. The language supports different platforms like windows,mac and linux which makes it a highly robust cross platform working language and all the challenges that one faces during development will chip away by these features of the language.

Similarly python percepts at an extremely efficient level. It provides us with libraries such as bokeh, seaborn and matplotlib which makes shocking representations.

Fig 3.1 Plotting Libraries

3.2.2 PANDAS

Pandas is a well known Python bundle for information science, and in light of current circumstances: it offers amazing, expressive and versatile data structures that make data control and assessment straightforward, among various things. The DataFrame is one of these designs. Pandas is an undeniable leveling and data controlling. It takes help from numpy-pack and the important data strucuture used for it is called as dataframe. The license which is used for storing as well as controlling you to even dataset which has conflicts or portions from elements.

It's based on top of the NumPy bundle, which means a great deal of the design of NumPy is utilized or reproduced in Pandas. Information in pandas is regularly used to take care of factual examination in SciPy, plotting capacities from Matplotlib, and AI calculations inScikit-learn.

3.2.3 JUPYTERNOTEBOOK

Jupyter Notebooks offer a decent climate for using pandas to do data examination and showing, anyway pandas can moreover be used in content devices basically. Jupyter Notebooks empower us to execute code in a particular cell instead of running the entire record. This saves a huge load of time when working with colossal datasets and complex changes. Diaries similarly give a straightforward technique to picture pandas' DataFrames and plots. In actuality, this article was made altogether in a JupyterNotebook.

There are two sorts of information structures in pandas: Series and DataFrames.

1.Series: a pandas Series is a one dimensional information structure ("a one dimensional ndarray") that can store esteems — and for each worth it holds an exceptional file, as well.

2.DataFrame: a pandas DataFrame is a two (or more) dimensional informationstructure – fundamentally a table with lines and sections. The sections have names and the lines have files.

The individuals who know about R realize the information outline as an approach to store information in rectangular frameworks that can without much of a stretch be outlined. Each line of these networks compares to estimations or upsides of an example, while every section is a vector containing information for a particular variable. This implies that an information casing's lines don't have to contain, however can contain, similar kind of qualities: they can be numeric, character, intelligent, and so forth

Presently, DataFrames in Python are practically the same: they accompany the Pandas library, and they are characterized as two-dimensional marked information structures with sections of possibly various sorts. When all is said in done, you could say that the Pandas DataFrame comprises of three primary segments: the information, the record, and thesections.

Right off the bat, the DataFrame can contain information that is:

- a Pandas Series: a one-dimensional named cluster equipped for holding any information type with pivot names or record. An illustration of a Series object is one section from aDataFrame.
- a NumPy ndarray, which can be a record ororganized
- a two-dimensionalndarray
- dictionaries of one-dimensional ndarray's, records, word references or Series.

3.2.4 NUMPY

Numpy is the center library for logical figuring in Python. It gives a superior multidimensional cluster item, and apparatuses for working with these exhibits. In the event that you are as of now acquainted with MATLAB, you may track down this instructional exercise valuable to begin with Numpy. A numpy bunch is a cross section of characteristics, the aggregate of a comparable kind, and is recorded by a tuple of nonnegative entire numbers. The amount of estimations is the situation of the group; the condition of a display is a tuple of numbers giving the size of the bunch along every estimation.

NumPy very much resembles sci-py, sci-kit learn, pandas and is amongst the groups which cannot be overlooked while working on or studying information sciences. Primarily on the grounds that this library gives you a cluster information structure that holds a few advantages over Python records, for example, being more minimized, quicker access in perusing and composing things, being more helpful and more effective.

It is a library under the python project that defines the central collection for logical processing within the python framework. Apart from this it has an assortment for devices along with procedures which could be utilized for settling over a PC numerical algorithms involving issues from Engineering as well as sciences fields. Amongst apparatuses one of them contains an elite multi dimensional clusters having an amazing information structured on effective calculation for exhibiting as well as networks.

Within the exhibits, there is large measure for undeniable levels numerical capacities work for frameworks as well as clusters.

The arrays define a cluster of pointers that are arranged in a linear sequence, they have every cell has its own address in memory, data and datatype :

• By default the first pointer indicates the address of the first data inside thearray.

• Data types define what kind of data is stored in each cell of anarray.

• The shape has basically to do with the size of thearray.

NumPy works on multivariate exhibits. We'll jump into the entirety of the potential kinds of multidimensional exhibits later on, however until further notice, we'll center around 2-dimensional clusters. A 2-dimensional cluster is otherwise called a framework, and is something you ought to be acquainted with. Truth be told, it's simply an alternate perspective about a rundown of records. A network has lines and sections. By determining a line number and a section number, we're ready to extricate a component from a network.

Numpy cluster can be utilized by the function called as numpy.arr. On the off chance that we pass in a rundown of records, it will naturally make a NumPy cluster with similar number of lines and segments. Since we need the entirety of the components in the cluster to be glide components for simple calculation, we'll leave off the header column, which contains strings. One of the impediments of NumPy is that every one of the components in a cluster must be of a similar kind, so on the off chance that we incorporate the header line, every one of the components in the exhibit will be perused in as strings. Since we need to have the option to do calculations like track down the normal nature of the wines, we need the components to all be drifts.

3.2.5 MATPLOTLIB

Plot for information is broadly conceivable for intelligent manner with the use of plotting library called matplotlib. The Plot of diagrams has information visualizs, This trait is accomplished through utilization of Matplotlib-plotting library. It utilizes mostly utilizes many universally useful Graphical User Interface tool compartments,

like uiPython, Tknter, doe, and so on, for giving objective-situated API to inserting plotting within applications.

John Tracker initially composed Matplotlib-library, The engineer for it was M. Droettboom. It is an openly available py library fundamentally utilized in specialization as well as logical processing. It generally utilizes for python-SciPy and its logical estimation requires plot in charts or graphs. It has a library like GNUplot. Principle benefit towards GNU plot has matplotlib. The interest in py is the prominence prominence consistently increasing for matplotlib.

Apart from this justifying also includes the option in contrast to MATLAB, in the event that for its utilization at mix of Scipy as well as Numpy . Besides matlab being costly and not open source matplotlib has a free open sourced platform. Also, the objects -arranged get used like a article situated way. Along with this it tends for being broadly useful Graphical User Interface tool stash. Apart from this the procedural pylab intends to seen like MATLAB.

3.2.6 DJANGO

Django is a high-level Python Web framework that enables rapid event processing and a clean, straightforward design. Working with experienced engineers, it addresses a substantial portion of the difficulty of Web development, allowing you to focus on writing your application rather than rehashing a problem that has already been handled. It's open source and free.

- Django was created to assist designers in getting applications from concept to completion as quickly as possible.

- Django handles security properly and supports designers in avoiding a variety of common security blunders.

- Django's ability to expand quickly and elegantly is influenced by some of the busiest websites on the Internet.

CHAPTER 4
MODULE DESCRIPTION

4.1 INTRODUCTION

Our approach combines two distinct but connected ideologies. Natural Language Processing, the Emotion-Words Set, and a few abstract characteristics are all used in the main technique. It tries to collect and grade text based on the feelings expressed in it. To represent tweets, the following technique employs common classifiers such as SMO and J48. Finally, we combine both of these techniques to provide a Hybrid way for more successfully managing distinct feelings in the content. It's worth noting that, despite the fact that we've used the tweets model extensively throughout this research, this estimator is standard and can detect and analyse sentiments in any piece of text.

Term recurrence and reverse record recurrence are chosen on the highlights at that point to provide the feeling highlights a better score, and we apply some unique controlled calculation for feeling categorization from that point forward. In this work, automated AI methods called Support Vector Machines were utilised to classify feelings. We tried to understand ourselves by recognising a sensation in a book archive. In figure 1 below, we present our suggested system. Each exchange is shown in subtlety in the pieces that follow.

The SVM machine learning algorithm used to estimate the result to predict emotion of users using twitter API.

4.2 SYSTEMARCHITECTURE

We depicted honest Bayes classifiers as a set of representation methods based on the Support Vector Machine. We release as an example each pair of characteristics accumulated from each other using this single machine learning technique, which is surely a collection of algorithms each of which distributes a normal in vogue.

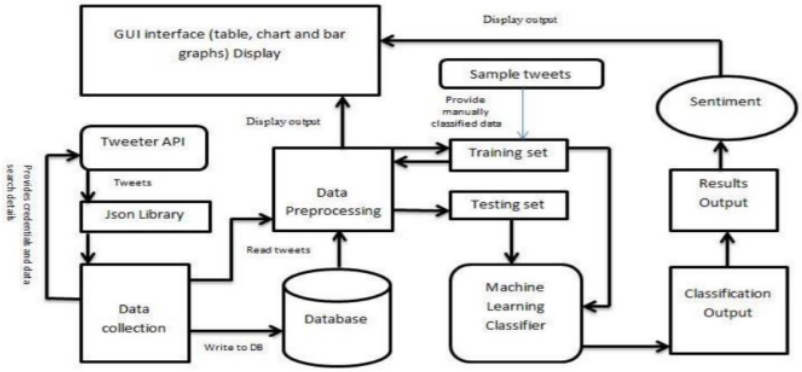


Fig 4.1 Architecture Diagram

As Shown in Figure 4.1, Initially data is loaded, Screening score, gender, and other attributes. To get insight into the data set, exploratory data analysis is done to know patterns and detect anomalies. Data Visualization is then applied to get the graphical representation of the data. Since it is a classic classification technique binary labels are applied to get the proper result. To remove the missing values in the data set data imputation techniques are applied to fill them. At last, the data is provided to different machine learning models and the best performing model is then used to predict user depression level as well as emotion from text.

4.3 TEXT PRE-PROCESSING:

We can go on to pre-handling and highlight extraction since we have preparing and test sets from the 60Users dataset.

A A method of transforming an enormous rundown of words into a more sensible request is standardization (otherwise called pre-preparing). At the point when we expect text to be in an exact organization before we can do anything with it, this is helpful. Standardization by space is a well known practice. For every application, there is nobody size-fits-all arrangement. In addition to other things, standardization may incorporate eliminating images and accentuation, supplanting characters, for example, $200 with the expression "200 dollars," and changing content over to lowercase. Emojis like :), :(are normally subbed by literary partners, like glad and sad, with regards to Twitter.

We eliminated hyperlinks, accentuation (with the exception of, and $), unique characters, and @BellLetsTalk) from the tweets in our application. We likewise change the instance of the content. At last, tweets with less than 5 words were wiped out. During the cleaning step, total tweets were killed, while the normalizing step disposed of characters from a tweet's substance. Stopwords are taken out in numerous applications, yet we excluded this progression to a great extent since we needed to safeguard words like "I," "me," and "you" (first individual pronouns). As per a survey of the writing, first individual pronouns are helpful pointers of gloom. Second, we needed to keep the content as comprehensible as could really be expected.

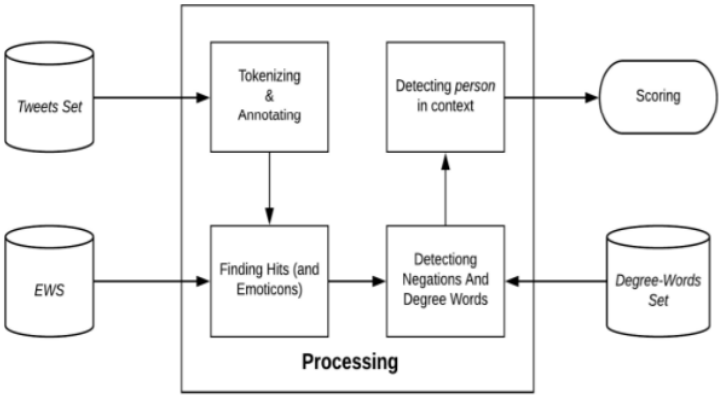


Fig 4.2 Preprocessing of Dataset

On applying missing value imputation, we found that ethnicity, relation, age were having missing values. It was found that ethinicity contained 95 total number of missing values. Whereas Relation contained 95 total number of missing values present in the relation column of our dataset. Apart from this the age attribute also contained missing values.

4.4 FEATURE SELECTION:

We by and by have text in the association we need after pre-planning. The extraction of features from text is the accompanying stage. Checks of Positive and Negative terms, counts of pity related terms, checks of first and second individual pronouns, and the Bag of Words were among the traits isolated.

Both the readiness and test sets have these qualities decided. Exactly when a model is set up with seven features, the test set ought to have comparable seven features. The limit words module returns four characteristics, the decline words module returns one, and the pronouns module brings two back. The proportion of features returned by the Bag of Words (BOW) module, on the other hand, isn't fixed.

4.4.1 POLARITY WORDS:

We should initially recognize which words are positive and which words are negative prior to tallying the quantity of extremity words, i.e., Positive and Negative words. The AFINN net's AFINN-111(Nielsen, 2011) word list is utilized for this. AFINN is an assortment of English words with valence evaluations going from 0 to 1. - 5 (negative) and +5 (positive) (positive). Finn is the person who marks the words by hand. Arup Nielsen is a Danish plan firm. There are 2,477 words and expressions in the AFINN-111. We extricate four attributes from AFINN-111. vNegativeTermsvNegativeTermsvNegativeTerm (score - 5, - 4), PositiveTerms (scoring 1, 2, 3), NegativeTerms (scoring - 3, - 2, - 1), and vPositiveTerms (scoring 1, 2, 3). (score 4, 5).

Tokenizing documents/tweets is the initial step in counting polarity words. Tokenization is the process of breaking down a statement into words. Tokens are separated in our implementation by space, punctuation, and numerals. The statement "Been awake since 6 a.m. for no reason" comprises 8 tokens, for example. We may simply count the polarity terms among these tokens and save them as feature counts once we have tokens.

Table 2 Sentence converted to tokens

sentence	tokens
“been awake since 6am for no reason”	“been”, “awake”, “since”, “6”, “am”, “for”, “no”, “reason”

4.4.2 DEPRESSION WORDS:

We have to determine which keywords to count in order to count depressive words. We utilise a list of 204 depression-related words (see Table 5.5) from (Maigrot et al., 2016). The amount of depressive phrases in a tweet was assessed as a feature. We count all depression words as a single characteristic rather than calculating the frequencies of each phrase.

4.4.3 PRONOUNS:

The first and second person pronouns are counted separately as characteristics termed "firstPronounCount" and "secondPronounCount," respectively. “I,” “me,” “my,” “myself,” “mine,” “we,” “our,” “ours,” “ourselves,” and “us” are all first-person pronouns. The pronouns "you," "your," "yours," "yourself," and "yourselves" are all second person pronouns.

Fig 4.3 Depression words

"agonized","aid","alienated","alienation","alone","anger","angry","anguish",
"anguished","antidepressant", "anxiety","anxious", "attempt","awful",
"barren","beaten","better place without me", "blas","bleak","bleeding", "blue",
"child","communication","concern","confusion","courage","crestfallen","cruel",
"crushed","crying", "cycle","dead","death","Death-seeking","debilitating","defeated",
"dejected","demoralized","depressed", "depression","descent","desolate",
"despair","despondent","detriment","devalued","devastated","die", "disappointed",
"discouraged","discrimination","disease","disinterest","disinterested","dismal",
"disorder", "dispirited","distracted","distressed","doctor","dog days","done with life",
"doomed","down","downcast", "downhearted","drained","drugs","effect","empty",
"endure","esteem","family","fatalistic","fatigued","fear", "fed up","feelings","fight",
"finality","friends","gain","gloomy","glum","grief","grieved","grieving","grim",
"hard work","health","help me","helpless","hopeless","hopelessness","Hot-line","hurt",
"I cry","I'm crying", "I'm done","immune","improvement","in despair","inability",
"inactivity","indifferent","insecure", "interested","involvement","irritable","isolated",
"isolation","joyless","kill","kill myself","lack", "life quality","lost",
"media","medication","medicine","melancholia","melancholy","mental","miserable",
"misunderstanding","moody","morose","necessary","need","negative",
"not pretty","nothing","option", "overcome","pain","panic","parents","passionless",
"patience","patient","pattern","pay attention","peers", "pessimistic","pills",
"pleasureless","prescription","prevent","prevention","progress","protect","quantity",
"reality","reckless","regretful","requirement","rotten","sadness","scared","security",
"Self-destructive", "separation","seriousness","signs","skills","solitary","somber",
"sorrowful","struggle","studies","succor", "suffer","suicide",
"sullen","sympathetic","symptoms",
"tearful","teenagers","terrified","therapy", "thoughts","tired","tormented","torture",
"tragedy","tragic","treat","treatment","trouble","troubled",
"uncertain","uncomfortable","unfulfilled","unhappy","unique","upset",
"victim","warning","weepy", "woeful","worried","worry","worthless","zero"

4.4.4 BAGS OF WORDS:

A word recurrence grid is all that Bag Of Words is. Pack of words doesn't return a fixed measure of attributes, not at all like extremity words, misery words, and pronouns, which all return a foreordained measure of qualities.

In the event that we figure BOW highlights for the preparation and test sets autonomously, highlights may veer, and the prepared model will be not able to anticipate on the test set. To bypass this, we register BOW qualities altogether from the preparation set. Accept the preparation set has 100 qualities. We just figure esteems for the 100 ascribes that are provided in the test set. This guarantees that the prepared model can precisely expect any new test information.

4.5 PREDICTION MODEL SELECTION:

There are a few distinctive Machine Learning calculations being used today. Backing Vector Machine (SVM), Naive Bayes (NB), Random Forest (or other tree-based calculations), and outfit approaches are the absolute most broadly used calculations in Natural Language Processing occupations. There is anything but a solitary calculation that can deal with all undertakings. Scientists regularly test various calculations prior to streamlining them for a particular assignment.

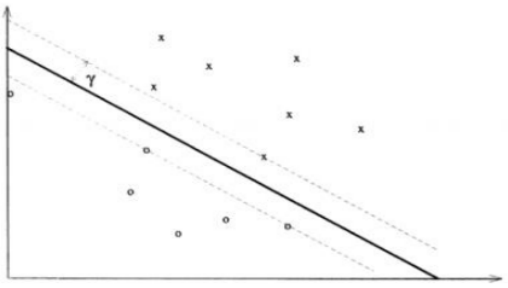
Discovering appropriate highlights is a significant piece of any Machine Learning measure. The calculations that group these attributes at that point search for designs. These properties are regularly gotten from text in NLP errands. n-grams and Bags of Words are two instances of highlights that depend on word frequencies. Different viewpoints, for example, an archive's feeling esteem, tone, clarity level, etc, are more issue explicit. The objective of producing highlights is to extricate data from text and convert it into a portrayal that an AI calculation can comprehend.

4.5.1 SUPPORT VECTOR MACHINES:

SVM represents Support Vector Machines, and it's an order calculation. SVM's preparation calculation creates a model from an assortment of named preparing models for a twofold class issue that recognizes an ideal hyperplane separating models from the two classes. It upgrades the edge, or the distance between the isolating hyperplane and the closest preparing occurrences.

SVM produces an estimate contingent upon which side of the hyperplane the example falls on when given a test.

Fig 4.4 Support Vector Machine



We have utilized Support Vector Machine in our model. We have picked a lot of seed words, including ordinarily used Emotion-words and emoticons from the EWS, consistently appropriated over all of the Emotion-Categories. We by then inquiry Tweepy using these seed words to develop an enormous data base of around 25,000 tweets. The seed words are used to ensure that we get tweets that express at any rate one of the six sentiments. The retweets are removed out to keep from any emphasis of tweets. The even allocation of seed words ensures that we get even dispersal of tweets over all of the Emotion-Categories, so our classifier isn't uneven. Regardless, we have analyzed that there is an enormous degree of tweets on Twitter imparting bliss, so we have kept the quantity of seed words having the Emotion-Category HAPPINESS, on the higher side. SVMs accomplish superior text arrangement, however they acknowledge high dimensional element spaces and scanty component vectors. Additionally, text identification uses SVMs which is strong to exceptions and doesn't need any boundary tuning. It finds a most extreme edge isolating hyperplane among2 classes of data. For different class SVM enhance the edge for one versus all classes of data., Linear SVM of the SK-learn pack are utilized. We introduced that straight pieces-based SVM plays out significantly better result.

4.5.2 LINEAR REGRESSION:

Linear regression algorithms are a type of ML algorithm in which we train a model to predict the behaviour of information based on certain factors. Because of direct regression, as the name implies, the two factors on the x- and y-axes must be related.

In that situation, a model may believe you are in business and anticipate a certain level of customer check that will be increased in this situation. Are you attempting to plan it, or are you attempting to determine what number will be tested? This means which my recent status will provide you with a more advanced planning strategy for the number of slower prices you require, or the number of representatives you require to serve the client. The idea here is to appraise the upcoming future worth based on authentic information by incorporating behaviour or examples from verifiable data. Sometimes the value will be linearly up, which means that when X grows, Y grows as well, or the other way around, which means they have a connection or there may be a direct descending relationship.

One model might be that the police department is on a mission to reduce the number of thefts; in this case, the diagram will be linearly descending. The straight relapse method is used to forecast a quantitative response Y from the indicator variable. A linear regression model can be represented mathematically as follows:

$$y = a + bx$$

Wherein a and b denoted by the equations:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

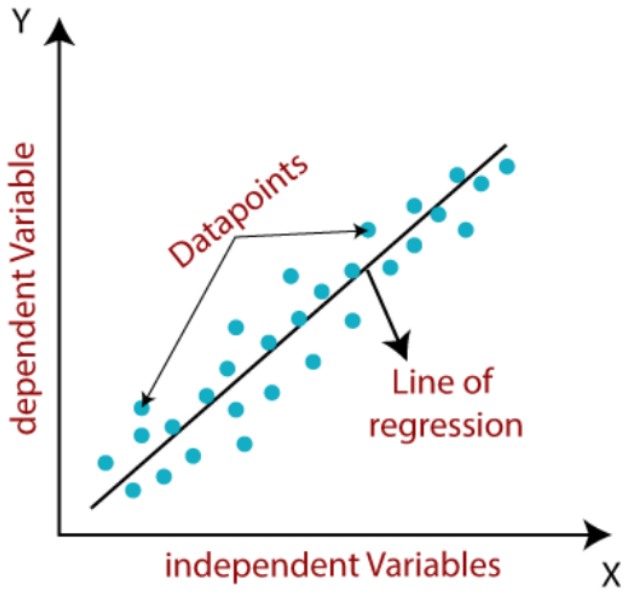
x and y are two variables on the regression line in this case.

The slope of the line is denoted by b.

a is the line's y-intercept.

x is a dataset independent variable, and y is a dataset dependent variable.

Fig 4.5 Linear Regression



4.6 CHOOSING BEST MODEL

It is necessary to measure the performance of a classifier in order to validate its accuracy and applicability. The most often utilised measures in the past have been “accuracy" and "error rate." The percentage of properly categorised cases is called accuracy, whereas the proportion of wrongly categorised occurrences is called error rate.

Let Actual indeed, Actual no be certifiable positive and negative class marks, and Anticipated indeed, Predicted no be anticipated positive and negative class names in a fundamental two-class order issue. A disarray grid () would then be able to be utilized to fabricate a portrayal of arrangement execution.

Accuracy = (T P + T N)/(Pc + Nc)

ErrorRate = 1 – Accuracy

Table 4.1 Best model column counts

	Actual yes	**Actual no**
Predicted yes	True Postive (TP)	False Positive (FP)
Predicted no	False Negative (FN)	True Negative (TN)
column counts	P_c	N_c

Because accuracy and error rate are inversely related, a classifier that has high accuracy and low error rate is regarded adequate for a job.

The purpose of training a user-level classifier is to predict if a user is depressed or at risk of becoming depressed. We begin by integrating data for each user in order to classify them at the user level. Three things are taken care of by the merging module. To begin with, it mixes the texts. all of the user's tweets into a single document Second, it summarises the characteristics determined during tweet-level categorization, i.e., the total number of tweets Counts of polarity words, depression words, and pronouns. It is also the third. utilises the tweet-level classifier's predictions to count the number of "depressedTweetCount" is a new feature that counts distressed tweets.

CHAPTER 5
SYSTEM IMPLEMENTATION

5.1 INTRODUCTION

Initially data is loaded, Screening score, gender, and other attributes. To get insight into the data set, exploratory data analysis is done to know patterns and detect anomalies. Data Visualization is then applied to get the graphical representation of the data. Since it is a classic classification technique binary labels are applied to get the proper result. To remove the missing values in the data set data imputation techniques are applied to fill them. At last, the data is provided to different machine learning models and the best performing model is then used to predict the user emotion and sentiment.Show the chances of a specific patient by utilizing the expectations made so that the forecast will beexact.

5.2 OVERVIEW OF PLATFORM

Anaconda Navigator is a workstation graphical user interface (GUI) for Anaconda distribution that allows you to dispatch programmes and manage conda bundles, conditions, and channels without having to use order line orders. Anaconda Cloud or a local Anaconda Repository are both good places to seek for packages. It runs on Windows, Mac OS X, and Linux.

Several logical bundles rely on explicit variations of other bundles to execute. Information scientists commonly use a variety of adaptations of various bundles, as well as a variety of situations, to isolate these diverse versions.

The software for placing orders on the phone Conda serves as both a bundle supervisor and a climate director, assisting information scientists in ensuring that each bundle adaption has all of the circumstances it requires and functions correctly. Pilot is a point-and-click technique to working with bundles and conditions that eliminates the need to write conda orders in a terminal window. You may use it to find the packages you want, instal them in a climate, execute them, and update them all from inside Navigator.

An IDE regularly has a supervisor and a compiler window. There, we can compose our code, run it, and arrange it. Subsequently, it turns out to be a lot simpler to foster applications utilizing an IDE. The significant purposes behind utilizing an IDE are given beneath:

We can without much of a stretch compose our code in its content manager window. To store the asset records of a task, we have an undertaking manager window. It makes investigating of the code simpler. An IDE's autosuggest include helps us recorded as a hard copy proficient code. We can import different tasks for incorporating with the current one.

Without an IDE, an item web application computer programmer should pick, send, consolidate and manage the instruments freely. The integrated development environment gets gigantic quantities of changes and enhancements that are required for the whole framework, app and organization. This integrated toolkit gets expection expected of smoothing out coding progression which could recognize or restrict designing stumbles or syntactic blunders.

Jupyter is a not-revenue driven affiliation designed for "encouraging open-sourced coding, non licensed-rules, organizations in natural enlisting within distinct coding lingos". Project Jupyter maintains multiple vernaculars for many coding situations Jupyter is known for 3 place coding vernaculars maintained via Jupyter. Jupyter made as well as maintained instinctive enrolling things like Notebook, Hub and JupyterLab, which included front line variation of thesame.

It is an electronic shrewd for computation climate which makes Jupyterscratch cushion documents. Also "scratchpad" terminology conversationally pionts to various substances, by and large the Jupyterweb app, Jupyter-Python website trained professional and Jupyter record plan subject to settings. JupyterNB document gives JSON-formatted, followed by a molded creation, which contains organized once-over data/output units having code, text (using .md, math, plot and filled medias, by and large completing the ".ipynb" development.

5.3 IMPLEMENTATION DETAILS

5.3.1 INPUT PARAMETERS

The reason for preparing a client level classifier is to foresee if a client is discouraged or in danger of getting discouraged. We start by coordinating information for every client to order them at the client level. Three things are dealt with by the combining module. To start, it arranges the entirety of the client's tweets into a solitary report. Second, it includes the entirety of the factors decided during the tweet-level order, like the extremity, pity, and pronoun checks. Third, as another element called "depressed Tweet Count," it use the forecasts from the tweet-level classifier to count the quantity of upset tweets. Re computing highlights is equivalent to joining highlights. Two new components have been added to "depressed Tweet Count," specifically "Total Tweet Count" and "depressed Tweets Percentage." Total Tweet Count is a numeric variable that shows the number of tweets someone in particular has in the dataset. The variable Depressed Tweets Percentage is a number.

5.3.2 CODING

We are importing the inbuild python libraries which will be needing to build the model. We will be making use of pandas and numpy library for generating the dataframe using which we will be using to make our Model. For data visualization we are using seaborn and matplotlib library. For the model building we used sklearn library which contains many in build algorithms.

We have included the needed libraries as per our requirement for the code, the code has been implemented with the help different algorithms like knn, lda and logistic regression. The best performance algorithm is chosen. The best algorithm is chosen on the basis output accuracy on the parameter of categorical analysis. The categorical analysis is done with the help of label binarizer.

For preprocessing of data we will first have to clean our dataset and remove the null values from the dataset and then convert the object types into string. After this we will be converting yes and no values into binary values respectively. By doing this preprocessing work we were able to train our model in the effective model.

Now we will be cleaning our dataset using lower, strip functions. Now we plot the graphs and find the relations between different attributes of the dataset.For converting the Multiple values we will apply label binarizer for converting those values into binary form. We also did some analysis on the data that in present in the dataset. On applying missing value imputation, we found that ethnicity, relation, age were having missing values. It was found that ethinicity contained 95 total number of missing values. Whereas Relation contained 95 total number of missing values present in the relation column of our dataset. Apart from this the age attribute also contained missing values.

On doing analysis on the date we found that one attribute contained a significant outlier. This means it can hinder our model making.Since it will show extreme values then the mean, median and mode will get affected and show wrong values. This will hinder our model performance. So in order to get best performance out of our model we had to change that outlier value. Age attribute was found to be having one extreme value of 382. So to fix this we have to remove this value else it will act as a noise for ourdataset.

Using sklearn library we will dived the dataset into two parts. Our first part will be training dataset and the other part will be testing dataset. The majority of the data will be used for training, the remaining portion will be utilized to determine the outcome. The training part will have seventy percentage of the data and testing part will have thirty percentage so the ration will be seven into three.

Now we will be preparing our dataset for the model by dividing the set into training and testing dataset.For this our X sample will be containing all the features except the class decision value. Whereas the Y sample will contain the dataset consisting of the decision class. Further we divided our dataset into 3:7 ratio for training and testing respectively. Then we applied different algorithms on them to find the best performing algorithm.

The algorithm which we applied is SVM (support vector machine). Since we had to do categorical data analysis, Logistic regression deals with categorical data analysis. After appling label binarization our data was converted into categorical data. Soo it was found that support vector machine was the best algorithm among the other algorithm including Naïve bayes and linear regression.

By applying this algorithm the performance measures have been found to be the best. With an accuracy of 94 percentage achieved, we can easily implement our code. The coding section has been included in the appendix.

5.3.3 IMPLEMENTATION SCREENSHOTS

```
In [20]: df_train.head()
```

Out[20]:

	Input	Emotion
0	i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so damned...	sadness
2	im grabbing a minute to post i feel greedy wrong	anger
3	i am ever feeling nostalgic about the fireplac...	love
4	i am feeling grouchy	anger

```
In [21]: df_train.tail()
```

Out[21]:

	Input	Emotion
15995	i just had a very brief time in the beanbag an...	sadness
15996	i am now turning and i feel pathetic that i am...	sadness
15997	i feel strong and good overall	joy
15998	i feel like this was such a rude comment and i...	anger
15999	i know a lot but i feel so stupid because i ca...	sadness

Fig 5.1 Head of the Dataset

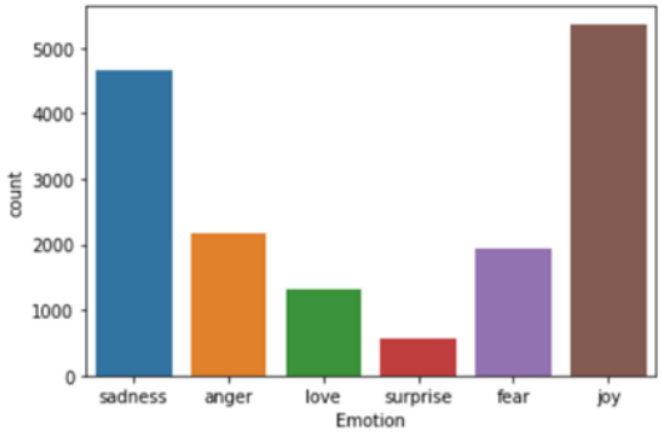


Fig 5.2 Emotion VS Count

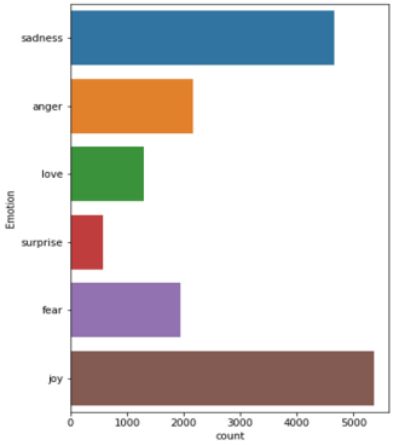


Fig 5.3 Count of Emotion

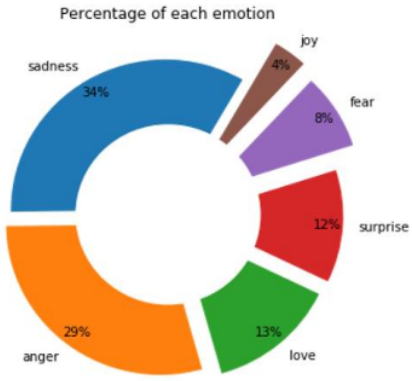


Fig 5.4 Percentage of each Emotion

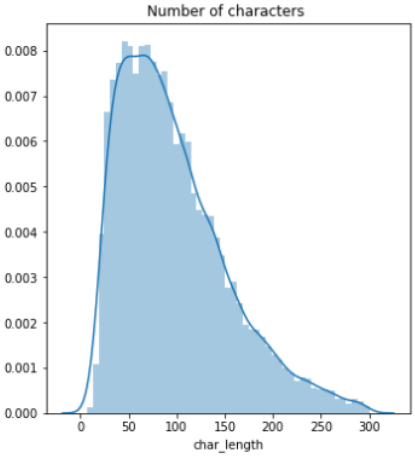


Fig 5.5 Number of characters

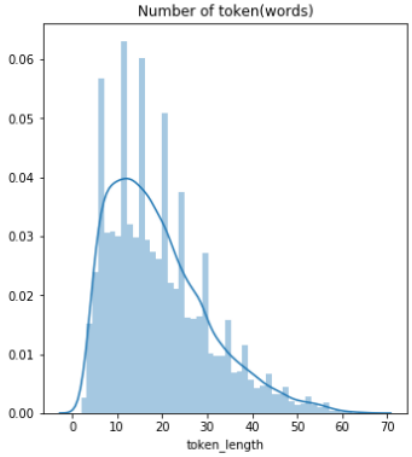


Fig 5.6 Number of tokens

```
plot_confusion_matrix(Y_test,Y_pred)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2223df8d7c8>
```

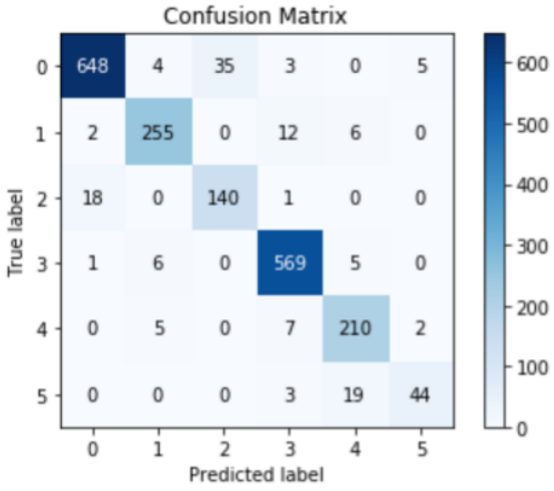


Fig 5.7 Confusion Matrix

```
In [65]: print(classification_report(Y_test,Y_pred))
```

	precision	recall	f1-score	support
0	0.97	0.93	0.95	695
1	0.94	0.93	0.94	275
2	0.80	0.88	0.84	159
3	0.96	0.98	0.97	581
4	0.88	0.94	0.91	224
5	0.86	0.67	0.75	66
accuracy			0.93	2000
macro avg	0.90	0.89	0.89	2000
weighted avg	0.93	0.93	0.93	2000

Fig 5.8 Performance of model

```
In [*]: predict(str(input('Enter a sentence : ')))
Enter a sentence :
```

```
In [68]: predict(str(input('Enter a sentence : ')))
Enter a sentence : I am broken and sad. Please help me. Feeling lonely
Out[68]: 'sadness'
```

Fig 5.9 Prediction of User Emotion

CHAPTER 6
RESULT ANALYSIS

The outcome is estimated as far as particularity, affectability, and precision by utilizing the disarray framework and order report. The outcome relies upon how exact the model is prepared. Estimating execution is vital to check how well an order model work to accomplish an objective. Execution assessment measurements are utilized to assess the viability and execution of the arrangement model on the test dataset. It is essential to pick the right measurements to assess the model presentation like disarray grid, exactness, particularity, affectability, and so forth Following equations are utilized to discover the presentation measurements:

Table 6.1 Elements of Confusion Matrix

		PREDICTED ASD VALUE	
Actual ASD Values		True Positive (TP)	False Positive (FP)
		False Negative (FN)	True Negative (TN)

$$\text{specificity} = \frac{\square\square}{\square\square + \square\square} \quad (1)$$

$$\text{sensitivity} = \frac{\square\square}{\square\square + \square\square} \quad (2)$$

$$\text{accuracy} = \frac{\square\square + \square\square}{\square\square + \square\square + \square\square + \square\square} \quad (3)$$

Table 6.2 Prec, recall f1-sc and support of proposed model

	Prec	recall	f1-sc	support
0	0.95	0.96	0.95	695
1	0.95	0.95	0.95	275
2	0.88	0.79	0.83	159
3	0.96	0.98	0.97	581
4	0.89	0.89	0.89	224
5	0.80	0.74	0.77	66
accuracy			0.94	2000
macro avg	0.91	0.89	0.90	2000
weighted avg	0.94	0.94	0.94	2000

Fig 6.1 Performance of Proposed Algorithm

After pre-dealing with and removing unwanted tweets for feeling word, we scrap and plan to test the dataset to get ready and play it on ordinary ML Classifiers. Backing Vector Machine (SVM) was applied on our model. We got an accuracy of about 94%. It was discovered that Support Vector Machine was exceptionally doable with higher precision than other different algorithms.

CHAPTER 7
CONCLUSION AND FUTURE WORK

Feeling identification is perhaps the hardest issue to settle. Recognizing feeling from text is testing work and the majority of the exploration works have some caring constraints in particular, language equivocalness, various feeling bearing content, text which doesn't contain any passionate words, and so forth However we have attempted a few ways to deal with recognize feeling from Twitter. In future, a system could be set up for thus invigorating the pack of-words which we made, in view of new tweets and data separated. Using our strategy, numerous captivating applications can be made, for instance, an extra to a long reach casual correspondence site showing the new perspective of all of your mates. Likewise, our investigation of Twitter can be reached out to the advancement of a constant framework, dissecting emotional episodes and feelings on Twitter.

The proposed models of feeling order in content in this paper can characterize text dependent on idle semantic investigation of things and action words in the sentences. We looked at the outcomes between two models, i.e. single word and single word joined with word. This can separate the feeling from the sentences which are then contrasted and the outcome named by the perusers. The delegate enthusiastic word is then used to group the feeling of the sentence utilizing the cosine similitude. We directed tests and the outcomes show that our methodology is promising. We dissect feelings from story messages dependent on the enthusiastic words addressing every story sentence. Thusly, our methodology doesn't consider the context oriented data that can range multiple sentences.

In view of this case, it very well may be reasoned that model averaging can be a decent option in the anticipating of reaction variable in the numeric scale as well as in the all out scale by executing the calculated relapse measure. The ASD information has irregularity classification in the reaction variable. The strategic model averaging technique has better exactness and affectability in the assessment expectation of a class of ASD endured. Future examination will zero in on adaptability and accelerating the all out reaction time to additionally improve the client experience.

Analysis and Derivation of Optimum Domain Specific Semantic Model for Detecting Depression Text From Twitter Stream

ORIGINALITY REPORT

4%	1%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ML Tlachac, Elke Rundensteiner. "Screening For Depression With Retrospectively Harvested Private Versus Public Text", IEEE Journal of Biomedical and Health Informatics, 2020 Publication	1%
2	Monica L. Tlachac, Elke Rundensteiner. "Screening for Depression with Retrospectively Harvested Private versus Public Text", IEEE Journal of Biomedical and Health Informatics, 2020 Publication	1%
3	Submitted to Gitam University Student Paper	<1%
4	www.ijert.org Internet Source	<1%
5	Submitted to University of East London Student Paper	<1%
6	developer.mozilla.org Internet Source	<1%

7	towardsdatascience.com Internet Source	<1%
8	Submitted to Coventry University Student Paper	<1%
9	www.aclweb.org Internet Source	<1%
10	www.ncbi.nlm.nih.gov Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Analysis and Derivation of Optimum Domain Specific Semantic Model for Detecting Depression Text From Twitter Stream

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36