Depression Detection using Tweets

A report submitted for the course named Project II (CS-300)

By

Chhavi Rana
Bachelor of Technology, VI Semester
Roll No. 19010113

UNDER THE GUIDANCE OF

Dr. Navanath Saharia
Department of Computer Science and Engineering

Department of Computer Science and Engineering
Indian Institute of Information Technology Manipur

May, 2022

Abstract

Depression is a common illness worldwide, with an estimated 3.8% of the population affected, including 5.0% among adults and 5.7% among adults older than 60 years . It is a significant public health burden. Despite being a relatively common mental illness, depression is often under-diagnosed, under-reported, and under-treated. One possible way to address this issue is through automatic screening of individuals for depression. This creates an opportunity to explore the possibility of automatically detecting individuals with depression through their behavior on Twitter.

Twitter has now become the most popular medium for self-expression and community building. More and more people are turning to it to express, write, vent or just be while being comfortable doing so because it's an accepted social norm. Using dataset extracted from Twitter of depressed and non-depressed users, this project aims at studying the characteristic features that make suitable indicators of depression in Twitter-Users, hence early detection of depression. The project has to put to application Machine Learning algorithms.

-

Declaration

I declare that this submission represents my idea in my own words and where others' idea or words have been included, I have adequately cited and referenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/sources in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from proper permission has not been taken when needed.

Chhavi Rana

Date: 28 April 2022 (19010113)

Department of Computer Science & Engineering
Indian Institute of Information Technology Manipur

Dr. Navanath Saharia
Assistant Professor

Email: nsaharia@iiitmanipur.ac.in

To Whom It May Concern

This is to certify that the report entitled **"Depression Detection using Tweets"** submitted to by Chhavi Rana, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree,diploma or a course.

Signature of Supervisor

(Dr. Navanath Saharia)

Acknowledgement

First and foremost, I offer my sincerest and deepest gratitude to my supervisor Dr. Navanath Saharia for his eminently capable supervision, fantastic feedback, creative ideas, gracious patience, constant encouragement and endless support. I feel truly blessed to have had such a hard-working mentor.

Secondly I would like to thank my parents and friends who helped me a lot in finalizing this project and completing it within the limited time frame.

- Chhavi Rana

Contents

List of Tables

List of Figures

List of abbreviations

	A
acc	Accuracy
API	Application programming interface
	F
FN	False Negative
FP	False Positive
	M
ML	Machine Learning
	N
NB	Naive Bayes
	R
RAM	Random Access Memory
	T
TF-IDF	Term Frequency- Inverse Document Frequency
TN	True Negative
TP	True Positive
	W
WHO	World Health Organizations

Chapter 1

Introduction

A human being can survive almost anything as long as she sees the end in sight. But depression is so insidious, and it compounds daily, that it's impossible to ever see the end.

– Elizabeth Wurtzel

Mental health is an integral part of health and well-being. Mental disorders can affect anyone, rich or poor, male or female, of any age or social group. Mental Health has long been ignored by everyone alike because of the lack of visibility in contrast to that of physical injuries or diseases.

This project focuses on the most common yet underestimated psychological illness which is a major contributor to the overall global burden of diseases: Depression .

About 280 million people in the world are suffering from/dealing with depression1 and about 90% of these don't know how to. The reasons being as trivial as lack of awareness, scarcity of mental health resources and the stigma attached to mental health issues. According to recent stats released by WHO, in October 2020 India was identified as one of the most depressed nations in the world, with about 7.5% of its population suffering from a serious mental disorder.

Depression if left undiagnosed is a progressive disease which in the later stage poses loss of functionality, retreating to isolating and at its worst poses a risk to a person's life and affects everybody else in his surroundings.

Meanwhile, the social media revolution introduced a new dimension in the world. People have adapted to it. It is the popular medium of self-expression and community building owing to comfort and efficiency it offers.

Inspired and motivated to help people realise the importance of mental health, the goal of our project is to exploit the massive data issued from twitter at identifying depression symptoms, so people can get help. By building an algorithm that can analyze tweets exhibiting self-assessed depressive features, it will be possible for individuals, parents, caregivers, and medical professionals to analyze tweets that signal deteriorating mental health far before traditional approaches currently do. Analyzing tweets allows for a low-profile assessment that can compliment traditional services and would allow for a much earlier awareness of depressive signs than traditional approaches.

1.1 Outline of the report

This report is organised around four main parts.

Chapter 2 describes Existing System Study

Chapter 3 describes System Analysis, Design and Implementation

Chapter 4 Conclusion and Future direction

1.1.1 Gantt chart

A Gantt chart is a graphical depiction of a project schedule. The Gantt Chart below shows a rough timeline followed to complete the project. Scheduling makes sure that we stay focused.

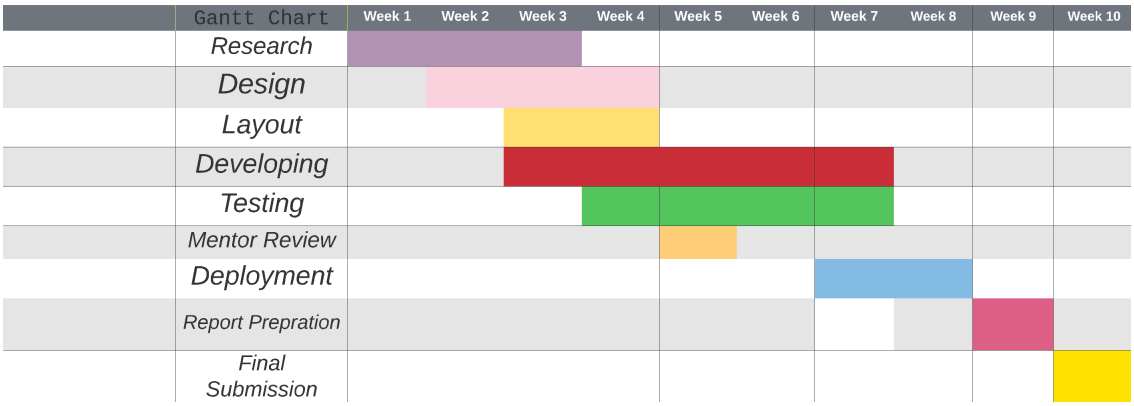

Figure 1.1: Gantt Chart

Chapter 2

Existing System Study

Outline: This chapter presents the following:

1. Introduction
2. Problem Statement
3. Proposed System

2.1 Introduction

Over the last few years, social media has been used to examine mental health by many researchers. There has been a lot of previous work done on analysis of Tweets using a variety of NLP techniques. Specifically a lot of these projects discussed the concept of generating embeddings for words and then feeding those embeddings into the model for training purposes.

2.1.1 Offline Depression Detection

Traditionally, depression detection is done in a face to face test with a clinical psychologist who bases her/his judgement on diagnostic criteria from manuals like the PHQ-9 i.e. the Patient Health Questionnaire. The guidelines and tests for clinical detection of depression check for verbal and behavioral patterns in the patient's physical world.

2.2 Problem Statement

About 10.6% of Indians have suffered from depression in the past 12 months, but more than half of them have not had it diagnosed. They may not have recognised it, may have feared the stigma of mental ill health, or had a sense of pessimism about treatment. So how can we better detect depression even if individual is either unaware or unwilling to share issues with professional.

2.2.1 Proposed Solution

The proposed solution is we can predict depression using user's tweet. Given the frequency with which Twitter is used by broad population, we can create dataset by scrapping tweets indicating depression. Tweets can be retrieved using the Twitter scraping tool TWINT using linguistic markers indicative of depression. Then we can construct our model using machine learning techniques.

Chapter 3

System Analysis, Design & Implementation

3.1 Introduction

Systems analysis is the process of observing systems for troubleshooting or development purposes. It is applied to information technology, where computer-based systems require defined analysis according to their makeup and design.

3.2 Hardware and Software Requirements

3.2.1 Hardware Requirements

1. Processor : Pentium 4 processor
2. Processor Speed : 2.4 GHz
3. RAM : 1GB
4. Storage Space : 40GB
5. Monitor : 1024x768 or 1280x1024

3.2.2 Hardware Requirements

1. IDE : Jupyter Notebook
2. operating system : Windows

3.3 Functional Requirements

3.3.1 Jupyter Notebook

Project Jupyter is a project and community whose goal is to "develop open-source software, openstandards, and services for interactive computing across dozens of programming languages". It was spun off from Ipython in 2014 by Fernando Pérez and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R, and also a homage to Galileo's notebooks recording the discovery of the moons of Jupiter. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is a NumFOCUS fiscally sponsored project.

Jupyter Notebook can connect to many kernels to allow programming in different languages. A Jupyter kernel is a program responsible for handling various types of requests (code execution, code completions, inspection), and providing a reply. Kernels talk to the other components of Jupyter using ZeroMQ, and thus can be on the same or remote machines. Unlike many other Notebook-like interfaces, in Jupyter, kernels are not aware that they are attached to a specific document, and can be connected to many clients at once. Usually kernels allow execution of only a single language, but there are a couple of exceptions. By default Jupyter Notebook ships with the IPython kernel. As of the 2.3 release (October 2014), there are 49 Jupyter-compatible kernels for many programming languages, including Python, R, Julia and Haskell.

3.3.2 Python

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a ”batteries included” language due to its comprehensive standard library.

Guido van Rossum began working on Python in the late 1980s, as a successor to the ABC programming language, and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features, such as list comprehensions and a cycle-detecting garbage collection system (in addition to reference counting). Python 3.0 was released in 2008 and was a major revision of the language that is not complete backward-compatible. Python 2 was discontinued with version 2.7.18 in 2020.

Python consistently ranks as one of the most popular programming languages.

3.4 Tools, Languages used

3.4.1 Tools

1. Numpy
2. Pandas
3. NLTK
4. sklearn
5. Scipy

3.4.2 Language

Python

3.4.3 Platform

Juypter Notebook or Google Colab

3.5 Theoritical View

3.5.1 Naive Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

The Naive Bayes algorithm is a theorem that work on conditional probabilities. That is, Naive Bayes calculate a probability event that will occur based on the calculation of the probability of events which has happened before.

$$P(A|\mathbf{B}) = \frac{P(\mathbf{B}|A)P(A)}{P(\mathbf{B})} \tag{3.1}$$

How to Apply:

1. Calculate the tf-idf weight per term from both classes (for binary classification) and create table for both depressed and non-depressed corpus.
2. The new input (sentence) that comes in is then tokenized to then calculate each the probability of both classes.
3. In the table, every word is required to have a tf-idf weight > 0 (there cannot be 0). For eliminating 0, performed Laplace Smoothing.

3.5.2 Laplace Smoothing

This is a technique for smoothing categorical data. A small-sample correction, or pseudo-count, will be incorporated in every probability estimate. Consequently, no probability will be zero. this is a way of regularizing Naive Bayes, and when the pseudo-count is zero, it is called Laplace smoothing.

3.5.3 tf–idf

Tf–idf short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today. A survey conducted in 2015 showed that 83% of text-based recommender systems in digital libraries use tf–idf.

3.6 System Design

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human machine interfaces, detailed design, processing logic, and external interfaces. Systems design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development.

3.6.1 Architecture

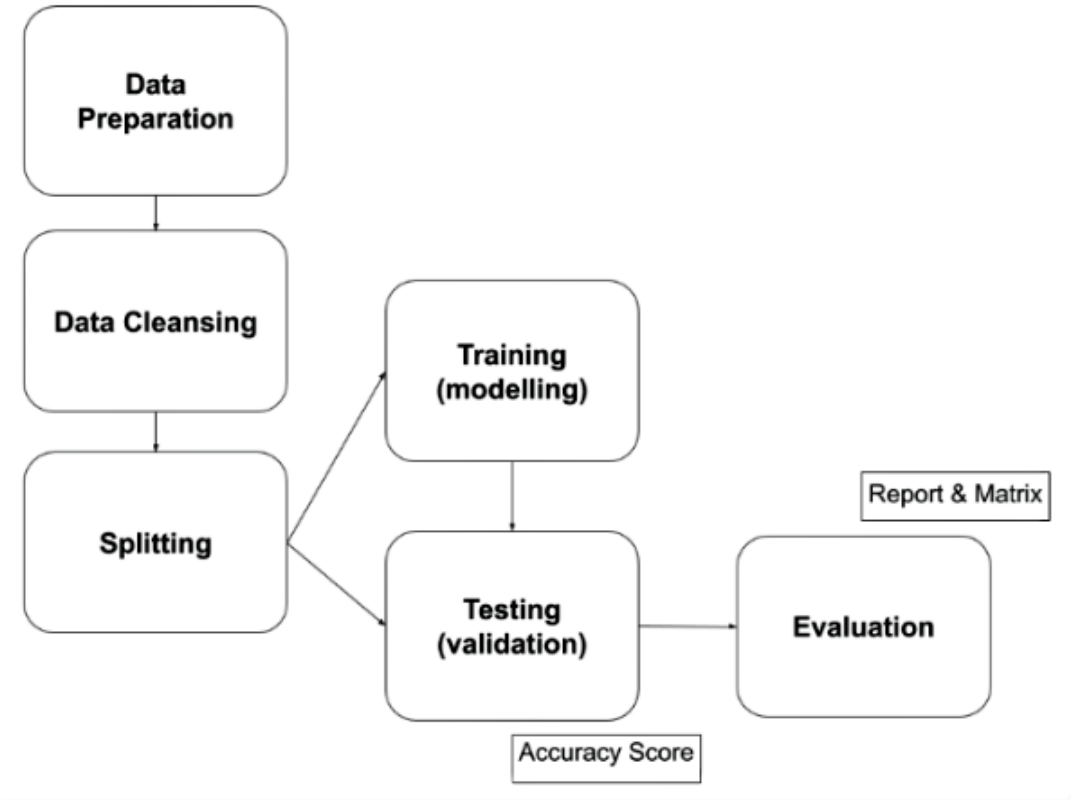

Figure 3.1: Architecture of Project

3.6.2 Data Preparation

I didn't find good quality and labeled datasets containing depressive tweets. So using Twint, tweets indicating depression were retrieved using linguistic markers indicative of depression. The non-depressed tweets were taken from Sentiment 140. After cleaning the data by removing all the unnecessary tweets, We remove the tweets with URLs, tweets with emails, lowercase all text, remove punctuation signs. and concatenating depressive and non-depressive tweets, a single CVS file was generated.

3.6.3 Data Processing

1. Read dataset csv file
2. Remove stop words
3. Apply stemming and lemmatization
4. Convert a collection of raw documents to a matrix of TF-IDF features using TfidfVectorizer with ngram_range(1,2)
5. Compute tf-idf for entire corpus: depressed and non-depressed corpus.

3.6.4 Model Classification

After the data has been processed, the data ready to be used to create a classification model using Naive Bayes. In this research, we divide (split) the dataset into two groups, namely training and testing. To create a model with classification performance a good one, good training data is needed many (variations) so that the Naive Bayes model able to make good predictions. We use 80% data on training and 20% data on testing.

3.6.4.1 Training

To train the classifier using training dataset, we have to compute following for entire corpus i.e. depressed and non-depressed Corpus.

- Compute Number of unique features
- Compute total sum tf-idf per term in corpus
- Compute total documents, non-depressed and depressed documents.

- Compute tf-idf weight per term

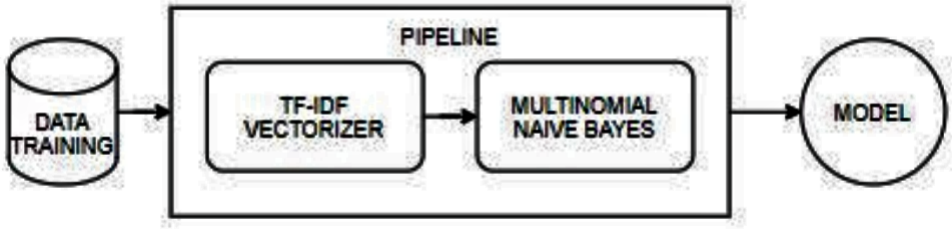

Figure 3.2: Pipeline

3.6.4.2 Testing

To run prediction on dataset csv file, read tweets from csv file, one row at a time and call naive bayes classifier to classify. Prediction is written to an output file. Predict function returns data frame that has new column prediction. Metrics function compares human prediction (label column) and system prediction column to find metrics.

Accuracy	0.90
Precision	0.88
Recall	0.91
F1 score	0.89

Table 3.1: Model evaluation

Chapter 4

Conclusion

We succeeded in making a model classification to predict depression based on text tweet, with accuracy (accuracy): 90.87%, sensitivity (recall): 89%, and precision (precision): 92.34%. The performance of our model was relatively high, most likely due to smaller size of dataset.There were several issue with the dataset that, if improved, would have made this project more practically applicable. The main issue with dataset was that gathered data from tweets was labelled on loose assumptions of Twitter account. It would have been nice to have a dataset of tweets from those who have been confirmed and known to have depression.

4.1 Future direction

1. Prepare a better and a large size data related to depressive tweets.
2. Attempting to mine information from other social networks. Dataset can be collected from other social networks such as Facebook, Tumblr, Reddit to establish a well-labelled dataset for analysing.
3. Cooperating with psychological industries for comprehensive research.
4. Making a Twitter Bot to detect Depressive tweets and connect the user to medical professionals.

Appendix A

Screenshot and Description of the Implemented System

A.1 Interactive Tool

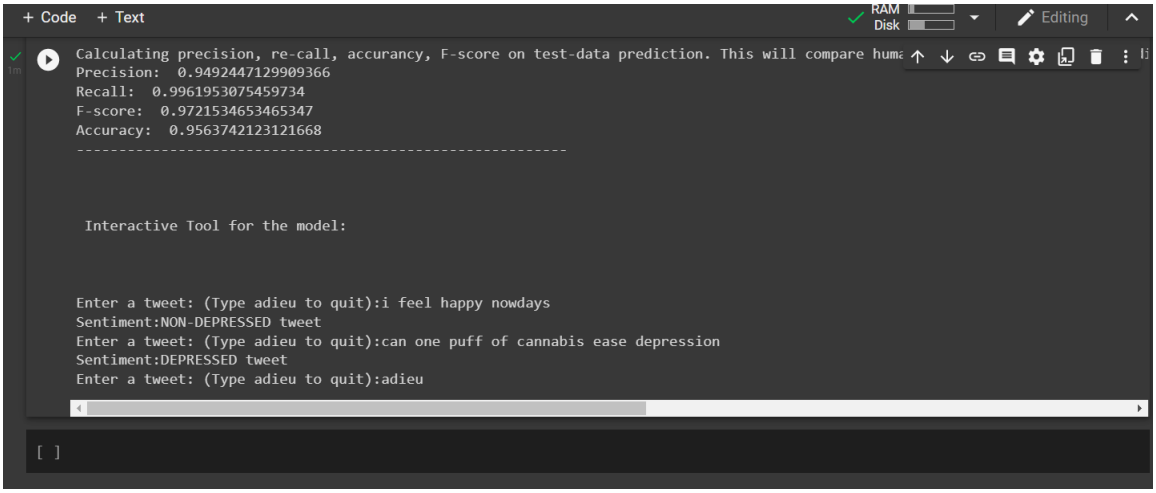

Figure A.1: Interactive tool for our model

We have a interactive tool for our model. We can write the tweet and check whether it is a depressed or non-depressed tweet.

Appendix B

User manual

B.1 Requirements

1. Jupyter Notebook

B.2 Step to install your implemented system

1. Download the zip file and extract the folder.
2. Launc Jupyter Notebook and navigate to the directory where you extracted the folder.
3. Open the depression-detector.ipynb notebook
4. Press Ctrl + Enter to run various cells in the notebook.

Bibliography

1. https://medium.com/syncedreview/applying-multinomial-naive-bayes-to-nlp-problems-a-practical-explanation-4f5271768ebf
2. https://medium.com/@rnbrown/more-nlp-with-sklearns-countvectorizer-add577a0b8c8
3. https://towardsdatascience.com/stemming-lemmatization-what-ba782b7c0bd8
4. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
5. https://en.wikipedia.org/wiki/Tf%E2%80%93idf