**Analysis of Suicidal and Depressing Posts based on Reddit**

Sanju Atwal  
*Master of Applied Computing*  
*University of Windsor*Windsor, ON  
[atwal3@uwindsor.ca](mailto:atwal3@uwindsor.ca) (110025865)

Naisargi Joshipura  
 *Master of Applied Computing*  
 *University of Windsor* Windsor, ON  
 [joshipu@uwindsor.ca](mailto:joshipu@uwindsor.ca) (110026949)

Satwinder Singh  
*Master of Applied Computing*  
*University of Windsor*Windsor, ON  
[singh9m@uwindsor.ca](mailto:singh9m@uwindsor.ca) (110033450)

Prabjot Singh  
 *Master of Applied Computing*  
 *University of windsor* Windsor, ON  
 [singh49@uwindsor.ca](mailto:singh49@uwindsor.ca) (110024705)

1. ABSTRACT

The central idea taken for the project by observing this online data is to fetch data from the famous networking website Reddit pre lockdown and post lockdown to analyze the deteriorating mental health of people, what exactly do they type and fetching those keywords to form our database and work on analyzing it.The team plans to fetch data from the keywords belonging to a category of tweets r/depression andr/suicide-watch on Reddit.The team attempts to train a model which can predict if a post or a tweet falls in these 2 categories and hence traces the records to depressive and suicidal tweets, collects the data and manages the database by analyzing and filtering tweets are required.

***Keywords: Reddit, Depression, Suicide, Database, Tweet, Post***

1. INTRODUCTION AND MOTIVATION

During the pandemic, people faced mental health issues and were struggling to cope up locked in home and hence, people started sharing feelings on social media either anonymously or becoming vocal. They sought social media like Facebook, Instagram, Pinterest, and Reddit saw such posts come to light. Reddit is a very popular social media app where people can post anonymously and share feelings. There are 2 sub reddits dedicated to depression and suicide. They are - r/depression and r/suicidewatch. They are specifically dedicated to helping people in these situations. Reddit has a good API with which we can post and fetch data. So, we decided to fetch data from reddit and perform predictive analysis using this data. The idea behind this was to bring forth data visualization and predictive data analysis as the key heroes in this project. Through analysing data with this model, the team attempts to train a model which can predict if a post or a tweet falls in these 2 categories and hence traces the records to depressive and suicidal tweets, collects the data and manages the database by analysing and filtering tweets are required. Data is scraped from these posts and further taken for data analysis. Aspects of data visualisation are also used. The focus is on Data Scraping and Big Data Analysis.

Motivation : The motivation for this project stems from the fact that people post a lot of content online and are very vocal about their mental health. In the lockdown phase worldwide in 2020, People started facing Depression and even became suicidal. Death of their beloved celebrities specifically due to suicide made things appear a lot more fragile than they were. Many people posted online and vented about they wanted to die or are feeling painfully low.

Many tweets surfaced and in countries like Canada, things seemed to be available pretty much online for the most part. Due to such a large data available and the lack of it being filtered out, the team decided to bring an idea into being regarding training a model that could predict if any post or tweet falls under the category of being either suicidal or depressing, so it could be filtered out and immediate help could also be made available through online medium for whoever posted so. This model serves to be helpful in predictive data analysis as well as lemmatization of important words in tweets and posts. Seeing such a large dataset gave the team motivation to work on the project and even think of future scope for the database, like trying to model Google’s State-of-the-Art NLP model BERT.

1. RELATED WORK

Existing works include some datasets collecting data for analysis on Kaggle which is a data science platform, but they only collect the data pre pandemic. The datasets in Kaggle have only some linkages to data during the pandemic. The focus of the team was specifically on data post pandemic and how people were affected even during the pandemic. The team found one research paper based on Deep learning and how it detects suicide risks from Facebook posts. But on Facebook, the data is not posted inconspicuously. The major drawback we found that because Facebook does not make any user anonymous, people were reluctant to share their feelings and hence, the dataset specifically for suicide and depression wasn’t collected. The team then thought of sourcing data from Reddit. Reddit is a social networking medium where people post data anonymously and share feelings. So, we targeted fetching data from reddit and were successful in the same. There is no other existent related work in this area, and this is a primitive, novel concept the team has worked on.

1. PROPOSED MODEL

The proposed model is a database using reddit API, posts belonging to which were collected from sub reddits r/depression and r/suicidewatch were collected. Also, to train the model with some normal posts, data was scraped and collected from API datasets like r/CasualConversations, r/DadJokes, r/Politics, r/GetMotivated and r/Television. Data scraping is a way to utilize data on the internet by fetching it into the system and embedding it into database. After collecting posts, we performed data analysis on r/depression and r/suicide watch. The data analysis included at first, cleaning the posts, for which the steps were:

1. removing any numerical data, symbolic characters or emojis from posts

ii) tokenizing the sentences into words – lemmatization of each word (simplifying each word and going by its meaning in the vocabulary) lemmatization replaces the word with its base form.

Eg: studies -> study, studying -> study, studied -> study

Playing -> play, played -> play, plays -> play

Wetted -> wet, wetter -> wet

Bowling -> bowl, bowled -> bowl

1. Stop word removal: we removed all the unnecessary words from our data which do not change the connotation of the sentence or overall meaning in general. The, a, an, so, what, when, therefore, et cetera… Re-joining the words to sentences and now we have a clean data. We selected only the relevant columns

[["title", "selftext", "author", "num\_comments", "is\_suicide","url"]]

The team worked to remove the null values in selftext column and then used WordCloud to plot the most frequently used words in r/depression posts, r/depression titles, r/suicidewatch posts and r/suicidewatch titles. The data analysis was performed using pandas and well as sql alchemy. SQL Alchemy is a python toolkit which enables the Jupyter notebook to run sql queries on data frames. Using sql queries, the team fetched the top authors who posted the most frequently and are categorised as the most vulnerable. then we analysed these top posts, the disturbing posts. The team performed data analysis to find the average length of titles and average length of actual posts and then, the average length of authors’ names and we found that average length of titles was about 39 characters, average length of posts was around 813 characters and average length of authors was about 12 characters. afterwards, we prepared the model for predictive analysis. We added a new column is\_normal = 0 for the posts belonging to suicide and depression category and we set it to 1 for other normal posts. Then we used count vectorizer to vectorize the textual posts and convert them into vectors which can be traced for prediction. Then we converted this data into training and testing data and then we used random forest classifier, which is used as Machine Learning classifier which is basically used for fitting 100 decision tree classifiers by default and averaging their prediction by default value. For our model, the default value of n\_estimators were kept 100, which is recommended. The max\_depth of the decision tree was set to NONE. This model gave us an accuracy of 0.8130 on testing data.

The RUC AUC score is the computation of area under the Receiver Operating Characteristic Curve (ROC AUC). It is a measure of the capability of model to distinguish between classes. The curve is plotted with True Positive Rate (TPR) against the False Positive Rate(FPR). TPR is plotted on the y-axis and FPR is on the x-axis. The value of TPR is calculated using a the formula:

TPR = TP/(TP + FN) and the value of FPR is calculated using the formula: FPR = FP/(TN + FP)

Here TP = True positives, which refers to the cases in which the predicted value of the model matches the actual value.

FP = False positives, where the model predicted a value other than the actual value.

TN = True negatives, where the model successfully predicted the negative values.

FN = False negatives, where the model predicted

positive for a negative value.

We got an RUC AUC score of 0.928.

1. DIAGRAM/RESULT

Chart

Description automatically generated

*Fig.1: Statistics showing Top Words used in r/depression Posts*

Chart

Description automatically generated

*Fig.3: Statistics showing Top Words in r/depression titles*

Chart

Description automatically generated

*Fig.2: Statistics showing Top Words used in r/Suicidewatch Posts*

Chart

Description automatically generated

*Fig.4: Statistics showing Top Words in r/Suicidewatch Titles*

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

1. LIMITATIONS/CHALLENGES

No project is achieved without facing obstacles. There were several challenges in this project. Data collection was a challenge because one seldom posts about depression or suicide, let alone anonymously or otherwise. Due to this, the team sought Reddit because users post anonymously there. The next task was to find normal posts along with finding posts on depression and suicide. From Reddit APIs, 4141 posts were selected, out of 1957 were null posts, and the rest were bifurcated into depressing and suicidal posts. Also, it was difficult for the analysing algorithm to analyse posts due to wrong spellings, use pf slangs, improper use of grammar in the posts. These were the challenges the team faced while developing the project and with efforts, were successful in developing the database.

1. CONCLUSION AND FUTURE WORK

We can see that there is a scope for improvement in our project as right now it guarantees 80% accuracy and we sense that this is an important area to work on, hence more development is needed. There is Google’s state of the art NLP model – BERT, the team can use this model to improve BERT’s accuracy, but we would need more data to use this model. We can make a GUI on which we can feed posts and detect the mood of the person based on the post. We can also develop a chatbot which can analyse live posts and help the ones in need by providing useful resources.

1. REFERENCES

* https://www.nature.com/articles/s41598-020-73917-0
* https://www.reddit.com/r/SuicideWatch
* https://www.reddit.com/r/depression
* https://www.researchgate.net/publication/339149662\_Suicidal\_Ideation\_Detection\_in\_Online\_Social\_Content