Detection of Suicidal Tendencies in Personal Reflections and Conversations: A Case Study of User Tweets

Damilare Samuel Adeola, 100374406

September 3, 2023

1 INTRODUCTION

It has been estimated that 4,912 people took their own life in 2020 in England(samaritan.org 2020). That means on a daily average, 13 people took their own life each day in 2020. The suicide rate was 10 per 100,000 people(samaritan.org 2020), and in figure 1, you would see that the amount of people committing suicide has been steadily increasing over the years.

Since the rate of suicide has been steadily increasing over the years, what have Primary Care Providers(PCP) been doing to mitigate this? What have the Clinical Psychiatrist and Psychologists been doing to plug this hole? The answer is not simple. Several researches have emerged whereby they sought to use Natural Language Processing(NLP) to create a model which can detect the likelihood of someone committing suicide from their text conversations at a fairly accurate rate and more importantly timely manner.

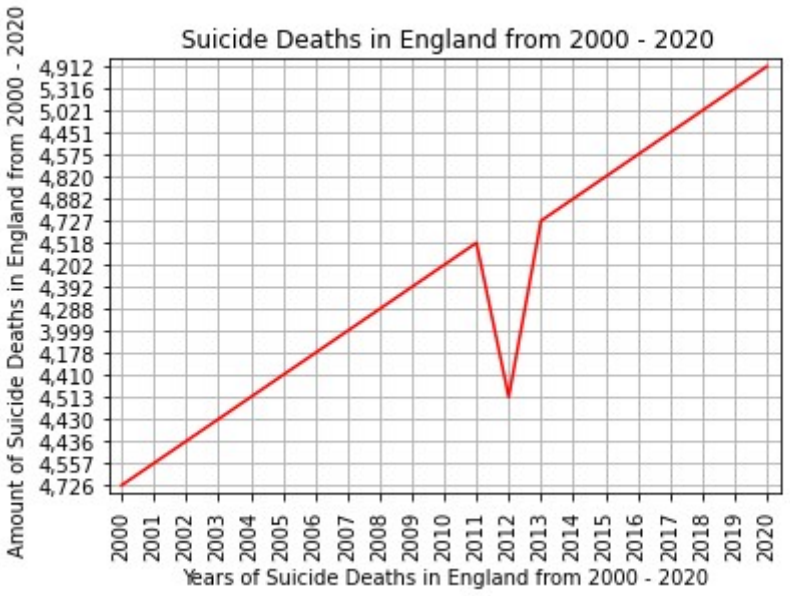

Figure 1: A plot showing the amount of people committing suicide annually in England from 2000 - 2020

Primary Care Providers, Psychiatrist and Psychologists actually do detect suicidal tendencies but they only detect these tendencies when patients with such suicidal tendencies come to them for personal consultation(Coppersmith et al. 2018). Hence, in the Health Care sector, there is a reactive solution to suicide prevention rather than a pro-active solution. And this is no fault of theirs as a PCP can only diagnose and treat what they are aware of. The use of Natural Language Processing(NLP) to be proactive in combating suicide prevention would give PCP and other Health Practitioners in the Mental Health Sector, the necessary tools to be proactive in not only treating people who have come to them with suicidal tendencies but to also identify people who have or are about to have suicidal tendencies.

The power to be able to predict via Social Media conversations and reflections, if a particular user has suicidal tendencies is what makes this approach effective. By using NLP and Machine Learning to perform Sentiment Analysis with Twitter data, an algorithm to detect suicidal tendencies in user tweets can be created so as to enable the relevant health stakeholders to help

such user and further save a life.

1.1 RELEVANCE OF THE STUDY

Digital Phenotyping can be defined as capturing and measuring the real-time features and quantification of human behaviour(Bidargaddi et al. 2017). With the amount of screen time spent by humans increasing every year, Fig 2 shows a horizontal bar chart of total screen time spent by people in different countries. The red bar is that of United Kingdom. It is estimated that in the UK, 50% of adults look at a computer screen for 4 hours or more each day and 26% of adults look at a smartphone screen for 6hours or more each day(Clayton & Clayton 2022). This means that 50% of adults spend 16.67% of their time each day looking at their smartphone screen. This statistic is given more context if we subtract the average sleep time from the 24hours so we can calculate the percentage of screen time as a proportion of when adults are awake. It is estimated that the average Briton gets just 6hours 19minutes sleep per night(Hall 2018). This means the average Briton is awake for 17hours, 18minutes per day. Now to get the proportion of screen time from the average total awake time:

Total Awake Time = 17.81

Screen Time = 4

Proportion of screen time in total awake time = (4/17.81)*100 = 22.41%

This means that the average Briton spends 22.41% of their total awake time looking at the screen.

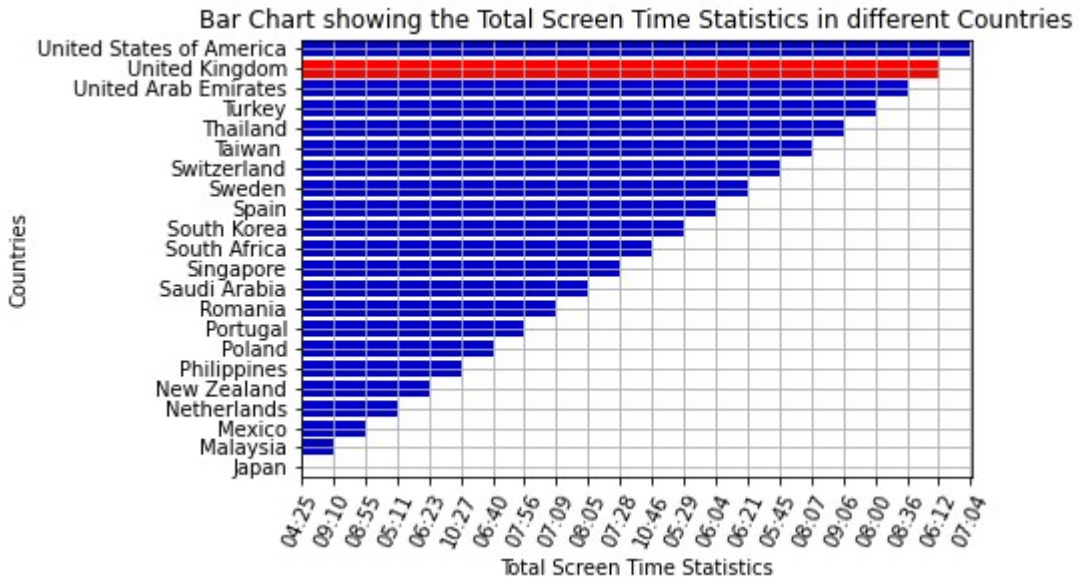

Figure 2: A bar chart of Total Screen Time by countries

What does this elaborate calculation mean to this study? It has great significance, what this means is that a lot of Digital Phenotyping is likely to happen among Britons, hence, we can actually use this data – digital conversations, reflections, tweets to be able to find out the suicidal tendency in a user's tweets since they spend a lot of time interacting in the digital world. Figure 3 shows a summary of the amount of screen time spent on social media.

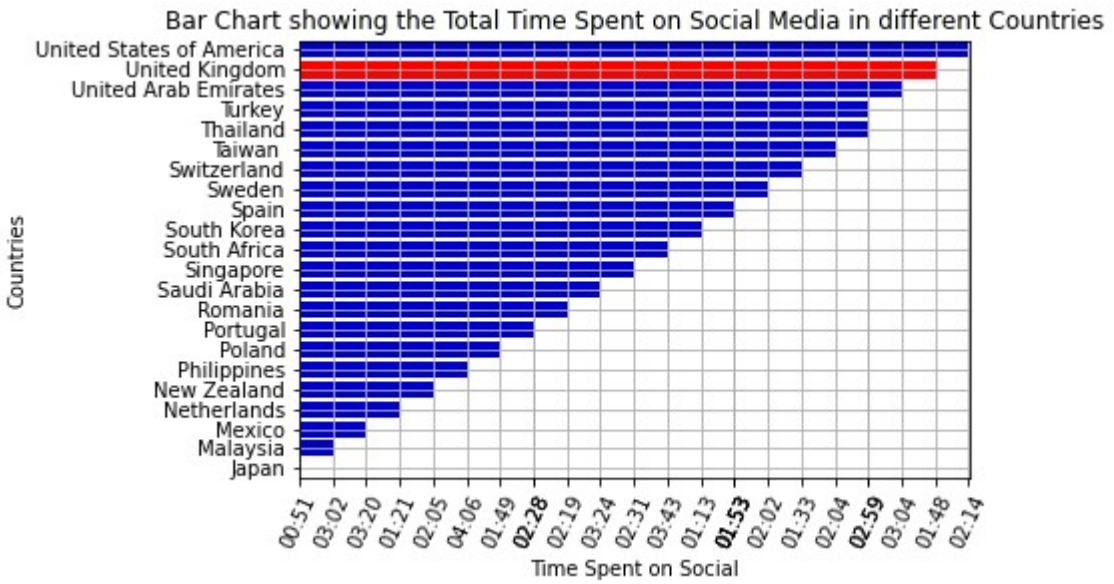

Figure 3: A bar chart of Total Social Media Screen Time by countries

1.2 AIMS AND OBJECTIVES

1. Build a Natural Language Model that would be able to predict the suicidal tendencies in a user's tweets. 2. Build a corpus of suicide text samples (this is not readily and easily available online). 3. Going further from the previous point, by building a corpora suicide text samples with categories and numerical scores for particular features. 4. Build a Natural Language Processing pipeline (Web Application) whereby a series of text can be passed through, and the NLP model would give a probabilistic estimate of whether that text contains suicidal thoughts.

2 LITERATURE REVIEW

A good amount of works have been published relating to using Natural Language Processing to detect and then classify users suicidal tendencies in Social Media. Apart from the increasing amount of screen time especially among teenagers to young adults(Boers et al. 2019), the deluge of information and live data has drawn researchers to using Social Media data - and this ranges from Facebook to Reddit to Twitter.

Twitter in particular has proven to be quite important in the study of mental health related issues of recent. Apart from the Samaritan 'Radar' scandal in 2014, there have been quite positive outcomes from using the Social Media's data in creating models for detecting suicidal tendencies. Twitter's text format and limited amount of characters make it a platform best suited for expressing one's thoughts and in use, that's how it feels, this is why it took a long time for Twitter to introduce the ability to delete tweets because really can we delete our thoughts?

These thoughts can also be suicidal, and Twitter has recognised it's role and put adequate services in place to help people(Coppersmith et al. 2018). But these services are not in real time. Research has been done whereby the data gotten from Twitter were from two groups: users who have tried committing suicide in the past and those who have not attempted to commit suicide(Coppersmith et al. 2018). This was done so the algorithm can be trained to be able to differentiate between those who are at risk of committing suicide and hence in urgent need of PCP and those who are not. Samples were also gotten of those who were a close match to having this suicidal thoughts but have not yet attempted suicide.

Two data sets were used in the experiments, one gotten from Twitter and the other from the now defunct ourdatahelps.org. When the two data sources were combined, they was a total of 547 users who have tried committing suicide and 418 users who actually committed suicide(Coppersmith et al. 2018). The categories of their data were fairly mixed, with most of the population comprising of females aged 18 to 24.

Despite the limited sample size, Deep Learning was used in trying to categorize the data. However, various layers were added to the layer to compensate for this shortcoming, like the use of the GloVe Embeddings(Glove stands for Global Vectors and it has to do with mapping the vectors of words to a global space(Pennington et al. 2014)). These various layers were introduced in order to prevent the model from overfitting with the training data. These word embeddings were later processed to make them better at capturing language relating to mental health(Coppersmith et al. 2018). The last layer involved using a 'sequences of word vectors which are processed via a bi-directional Long Short-Term Memory(LSTM) layer to capture contextual information between words'(Coppersmith et al. 2018).

There were two key results from their research, one was that the machine learning algorithms with high precision could delineate through text from social media those who would go on to attempt suicide and those who would not and secondly, that machine learning algorithms rely on a large amount of little 'clues' rather than whole phrases(Coppersmith et al. 2018).

Deep Learning always gives amazing results when it comes to classification problems, however, it comes at a high computational cost. Also, it stands to reason if for such a small sample size, if it was worth this cost. The results though tell otherwise, with their model having a true positive rate of 84% (Coppersmith et al. 2018).

3 RESEARCH METHODOLOGY

3.1 PROGRAMMING LANGUAGE AND PACKAGES

The following Programming Languages and packages would be used at different stages of the research:

1. Python
2. Natural Language Tool Kit(nltk)
3. pandas
4. numpy
5. matplotlib
6. wordcloud

3.2 DATA

CORPUS

A corpus can be loosely defined as a large collection of related text samples. For this research, we would be getting the corpus from Twitter. This would be done through Twitter's Developer's Application Programming Interface (API) for developers.

CORPORA

One of the main objectives of this research is also to create a corpora for suicidal tendencies. While searching online for relevant data for the project, it was tenuous to find any and those few organisations tasked with having such data had either been bought by bigger organisations or shut down (like ourdatahelps.org). This means there seems to be a gap in this area – text data for NLP tasks related to mental health in general and suicide in particular. Hence, from the corpus that would be derived from Twitter, a corpora would be created whereby each user tweet would be put into categories and numerical scores would be given to each user tweet – 0 for negative and 1 for positive or in this specific context – 0 for no suicidal tendencies and 1 for suicidal tendencies.

COLLOCATIONS and CONCORDANCE

Collocations can be defined as a series of words that appear frequently together in a given text. These series of words could be in pairs(Bigrams) or three(Trigrams) or four(Quadgrams). Collocations would help tremendously in the analysis as it would further the understanding of the Corpus and Corpora. Concordance on the other hand, is a collection of word locations along with their context.

WORDCLOUD

WordCloud are the plots and bar charts of NLP. There are used to extract more information from a given piece of text visually like the frequency of occurrences of different words. They help one to be able to visually perceive how often a word and in some cases the relationships or how often two or more words occur together in a given piece of text. Figure 4 below is an example of a wordcloud created from a collection of suicide notes.

Figure 4: An example of a word cloud. Data Source for text(russelljohn.net 2008)

3.3 DATA PREPROCESSING

In order to train the NLP model, we would need to feed it data that are relevant to the task – which is to be able to parse through a text and decide based on a probability score ranging from 0 to 1 if that text contains suicidal tendencies. Because of this reason, we would need to do data preprocessing so as to be able to collate relevant data and also to train the model adequately.

The data preprocessing would begin with first removing the stop words from the user tweets corpus. Stop words here refer to words that in the context of NLP, do not offer much meaning to our classification tasks. These words include prepositions, definite and indefinite articles etc like ‘the’, ‘a’, ‘an’ etc. These stop words can hamper the performance of our model and furthermore, increase the verbosity of each row data. The Natural Language ToolKit(NLTK) has a good method for this called:

$$stopwords = nltk.corpus.stopwords.words(\text{”}language\text{”})$$

3.4 TOKENIZATION

This can be defined as the segmentation of text into small, linguistic units such as words, punctuation, numbers, alphanumerics etc (Mikheev 2003). It can also be defined as the process of dividing words into separate units, called tokens and in the process discarding certain characters e.g punctuation. The usefulness of tokenization of text is that it helps when one wants to create a custom corpus and it also assists in inferring relationships between certain words or group of words in a text.

This phase of the research may involve two steps based on the result of the first step. In the first step, I would be using the NLTK token method:

$$tokens = nltk.word\_tokenize((corpus), width = (integer), compact = (Boolean))$$

Depending on the results gotten from the NLTK's token method, I may or may not apply further processing. It must also be noted that further processing may be done after this stage, if the token list contains punctuation. This can easily be removed using Pythons inbuilt method – isalpha().

3.5 FEATURE EXTRACTION

TF-IDF – TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY

TF-IDF is a popular algorithm used in extracting features from a corpus. It is a technique used to quantify words in a text document. It differs from the also popular Bag of Words(BOW) model in that while the BOW model measures the frequency of a word no matter the amount of times, which could lead to redundancy especially when it comes to nouns e.g a Guardian article about Amy Winehouse would definitely mention Amy Winehouse n-number of times; TF-IDF on the other hand measures the relevancy of words in the document.

TERM FREQUENCY

The acronym TF-IDF stands for Term Frequency – Inverse Document Frequency. Term Frequency(TF) has to deal with measuring how frequent a particular word occurs in a given corpus. This would inevitably depend on the length of the document. For example, a conjunction like 'and' would appear more in a document with 10,000words than in a document with 1,000words; what this means is that it is a good idea to normalise this frequency by dividing the TF by the total number of words. Hence we have:

$$TF(t, d) = countoftind/numberofwordsind$$

where:
t = term(word)
d = document(set of words)

The essence of Feature Extraction is due to the fact that computers cannot process human

language like humans do. In order for computers to be able to process language, they must be converted to matrices or vectors. This is one of the use-cases for Feature Extractions in NLP. However, in order to vectorize the words in a document, it is not enough to only consider the words that are only present in the document; we have to vectorize the list of all possible words in the document(Scott 2019). This process is known as Vocab.

DOCUMENT FREQUENCY

The document frequency measures the importance of documents in a set of corpus. One must recall that in NLP, a document is a sentence of human language. The formula for calculating the DF is:

$$DF(t) = occurrenceoftinNdocuments$$

where:
T = term(word)
N = count of corpus

The value of the DF just like the TF would also be normalised so as to keep it in the range of 0 to 1. This normalization is done by dividing the term by the total number of documents. The objective is to find out the relevance of the term and DF is the opposite of this(Scott 2019). It is for this reason the inverse of DF is calculated and used instead.

INVERSE DOCUMENT FREQUENCY

Inverse Document Frequency(IDF) is the inverse of Document Frequency. It is used to measure the relevance of a term in a document. When it is calculated, it gives a very low value to stop words (which you must recall has already been done before reaching this stage) and this gives us the outcome that is desired which is to have relative weights applied to all the terms in the document and most importantly for this weight not to be solely determined by their frequency but by their relevance and informativeness(Scott 2019).

IDF is gotten through the formula:

$$IDF(t) = N/df$$

where:
N = count of corpus
df = document frequency

3.6 CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

Three machine learning probabilistic algorithms would be used after the features for suicide tendencies have been extracted from the Twitter User Tweets corpus. They include Naïve Bayes Classification Model, Support Vector Machines(SVM) and the Logistic Regression Model for Classification.

The model with the highest accuracy and Precision score would be selected as the algorithm

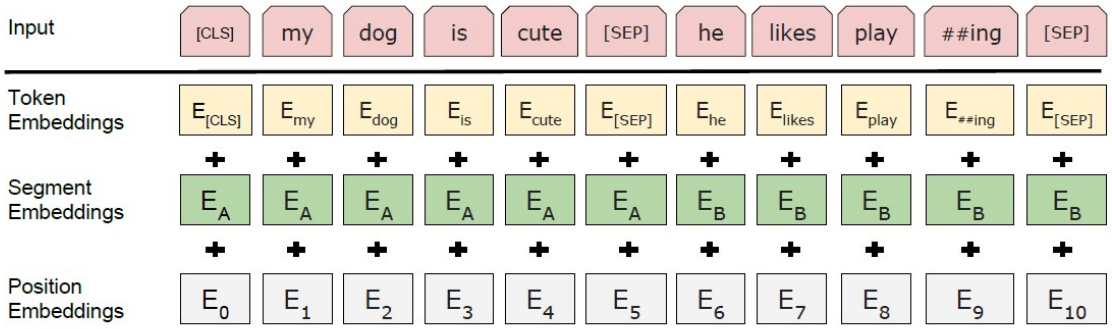

Figure 5: An image showing the BERT inputs (Devlin et al. 2018)

for classifying the NLP corpus.

3.7 THE BERT MODEL

BERT stands for Bidirectional Encoder Representations from Transformers and it involves using Transformers to train language models non-sequentially i.e. from left to right and right to left(Devlin et al. 2018). This is a departure from other NLP models which look at the text sequentially from either left to right or right to left. This makes the BERT model capable of context, an important trait in written human language. For most times in a sentence, the preceding part often informs the meaning of the ending part. e.g. 'I love you so much, I almost hate my life'. Using a sequential NLP model, it could take the second part of the sentence beginning with 'I almost hate my life' and think this must be a harmful signal because the model is unaware of the beginning part; hence lacking context.

This very reason, context, makes the BERT model appropriate for the NLP model that would be created to detect suicidal tendencies; as one would need a model that can derive as much context as possible from a User's Tweet in order to improve the classification performance, accuracy and precision.

3.8 MASKED LM(MLM)

In order to derive context from a sentence and move in a sentence non-sequentially, BERT relies on a method called Mask. Here, a part of the sentence is masked(hidden) from the model, then with a classification layer, BERT tries to predict the original value of the masked word. Usually, the amount of masked words could range from 10% to 15% etc with varying results(Devlin et al. 2018).

The prediction of the words has the following steps:

1. A classification layer is added on top of the encoder output.
2. The output vectors are then multiplied by the embedding matrix, transforming them into the vocabulary dimension(Horev 2018).
3. The probability of each word in this vocabulary is then calculated with softmax(Horev 2018).

4 ANALYSIS AND PROJECT MANAGEMENT

4.1 RESOURCES REQUIRED

For this research the main resources that would be required would be the corpus. This data would be gotten from the Twitter API development platform and as at the time of writing this proposal, the author can confirm that his application has been approved and he has been granted access to the Twitter API Developers backend. This means that the data for this project can be mined and preprocessing can begin as soon as possible. More details about the timeline are elaborated graphically in the Gantt Charts further down this section.

Depending on the amount of data mined, this research may also need high capacity GPU's to process the data. Especially since the BERT model would be used when creating the NLP model and it has as one of its drawbacks, computation costs though this can be waived due to the amount of context which the model provides.

Thirdly, in order to gain more domain knowledge and to use the resources available in the school, the author may collect primary data (where available) from people who have attempted suicide. This primary data would be added to the Corpora – which is one of the aims and objectives of this project. In the process of collecting this primary data, the author may collaborate with the Psychology department (a lecturer or student whomever is willing and available to assist) especially in the process of establishing:

1. More relevant search entries asides 'suicide' in the Twitter API backend.
2. Provide further context as to the causes of suicide which could in turn inform necessary decisions at various stages of building the model e.g if context is actually required when analysing suicide notes or a first person perspective on how to decipher a real suicide note from a fake one or prank.

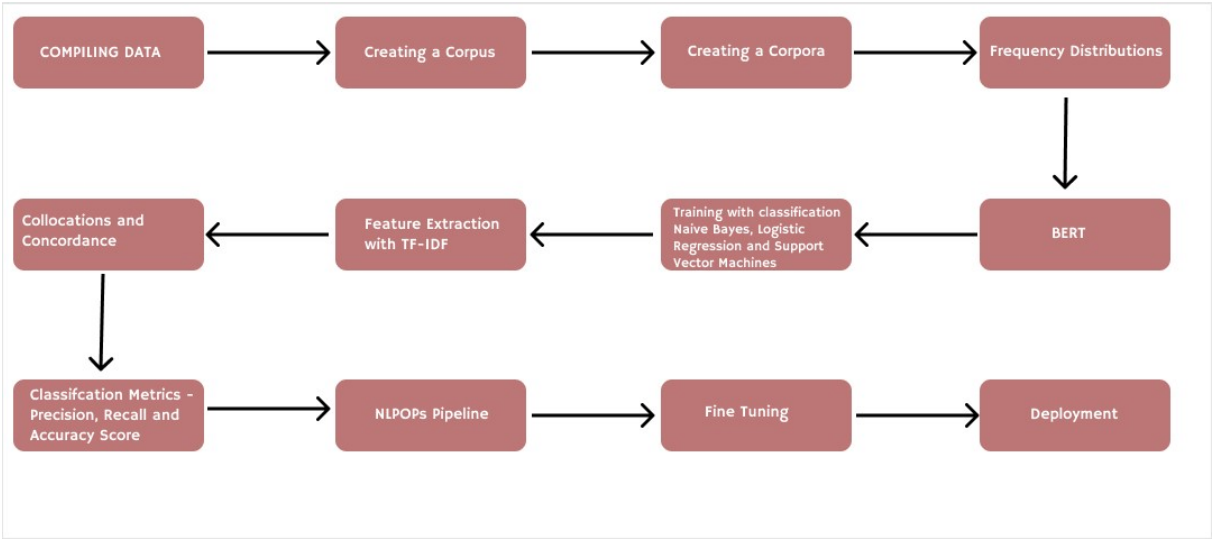

Figure 6: Flow Chart showing the different processes/stages involved in the research

4.2 RISK ANALYSIS AND MANAGEMENT

Figure 7 shows the Gantt Chart for the research, it shows the various start and expected end dates for each goal in the research. While figure 6 shows the Flow Chart of the various steps to be undertaken in the research. The Flow Chart is a good way to look at all the steps involved in a project sequentially.

ACTIONS	DURATION	START DATE	END DATE
Approval from Twitter	14 days	Mon 25/04/2022	Sat 07/05/2022
Begin Chapter 1 - Introduction	2 days	Tue 26/04/2022	Wed 28/04/2022
Finish Chapter 1	1 days	Wed 28/04/2022	Fri 29/04/2022
Review Data - Part I	2 days	Fri 27/05/2022	Sun 29/05/2022
Review Data - Part II	2 days	Sun 29/05/2022	Tue 31/05/2022
Creation of Corpus	2 days	Tue 31/05/2022	Thur 02/06/2022
Frequency Distributions of Corpus	1 day	Thur 02/06/2022	Fri 03/06/2022
Word Cloud of Corpus	1 day	Fri 03/06/2022	Sat 04/06/2022
Source for Primary Data	3 days	Sat 04/06/2022	Mon 07/06/2022
Interview Lecturer or Student in Psychology	2 days	Mon 07/06/2022	Wed 09/06/2022
Creation of Corpora	3 days	Wed 09/06/2022	Sat 12/06/2022
Cleaning and Refining Corpora	1 day	Sat 12/06/2022	Sun 13/06/2022
Literature Review - Further Research I	2 days	Sun 13/06/2022	Tue 15/06/2022
Literature Review - Further Research II	2 days	Tue 15/06/2022	Thur 17/06/2022
Finish Literature Review	1 days	Thur 17/06/2022	Fri 18/06/2022
Research Methodology	3 days	Fri 18/06/2022	Mon 21/06/2022
Preprocessing I	1 day	Mon 21/06/2022	Tue 22/06/2022
Preprocessing II - Stop Words	1 day	Tue 22/06/2022	Wed 23/06/2022
Preprocessing III - Stemming	1 day	Wed 23/06/2022	Thur 24/06/2022
Preprocessing IV - Lemmatization	1 day	Thur 24/06/2022	Fri 25/06/2022
Concordance & Collocation	1 day	Fri 25/06/2022	Sat 26/06/2022
Analysis - I	1 day	Sat 26/06/2022	Sun 27/06/2022
Analysis II - Sentiment Analysis	4 days	Sun 27/06/2022	Wed 01/07/2022
Analysis III - TF-IDF	2 days	Wed 01/07/2022	Fri 03/07/2022
Analysis IV - Classification	2 days	Fri 03/07/2022	Sun 05/07/2022
Analysis V - BERT	2 days	Sun 05/07/2022	Tue 07/07/2022
Deployment	2 days	Tue 07/07/2022	Thur 09/07/2022
Web Deployment	3 days	Thur 09/07/2022	Sun 12/07/2022
Summary & Conclusion	1 day	Sun 12/07/2022	Mon 13/07/2022
Revise & Review Entire Project	2 days	Mon 13/07/2022	Wed 15/07/2022
Target	1 day	Wed 15/07/2022	Thur 16/07/2022
Breathing Room	3 days	Thur 16/07/2022	Sun 19/07/2022
Final Submission	1 day	Sun 19/07/2022	Mon 20/07/2022

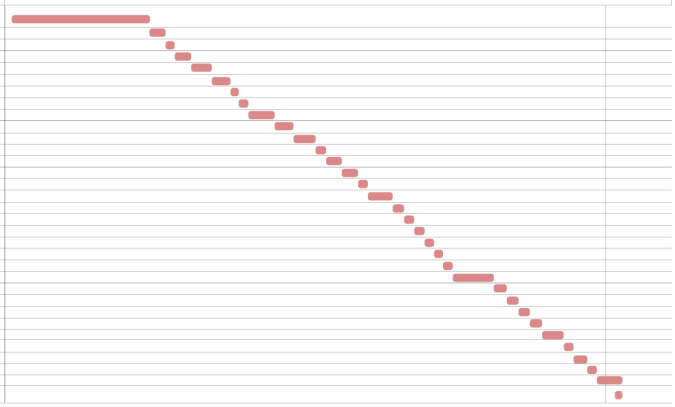

Figure 7: Gantt Chart showing the various project outcome deadlines

4.3 DEPLOYMENT

The NLP model would be deployed in a Web API. The Web API would be a simple interface whereby one can take a piece of text or sentences (that have been delimited) and then this text would be submitted via a HTML5 form. After the form has been submitted, the model then analyses the content at the backend, does the preprocessing and then finally gives a prediction. Finally, the client is taken to a page where the prediction is displayed to the user.

These deployment steps can be further simplified below:

1. User goes to a web page or address
2. User submits a form containing the text to be checked for suicidal tendency
3. Validations are done in the front end using HTML5 form validation attributes and elements
4. The server receives the validated form
5. The form passes through the preprocessing stage
6. (Optional) The outcome of the probability may be passed in a control flow statement whereby there would be different thresholds e.g a probability score of 0.89
would mean the document contains text with high suicide tendencies and that of 0.25 means the document contains text with low suicide tendencies
7. The result of the output is displayed back to the user.

The following tools and technologies would be used in the deployment stage:

1. HTML5 and CSS3
2. Javascript

3. Flask
4. Python
5. Jinja3
6. SQLite
7. Heroku
8. Git

4.4 LIMITATIONS AND ETHICS

The Ethics involved in this research could be a bit cumbersome especially with the poor historical precedent using Twitter's data in this area of research has caused in the past like the furore over SAMARITAN'S 'Radar' product(samaritan.org 2015) but unlike that product, ours does not invade any privacy as this research only seeks to know if a user's tweets has suicidal tendencies. It does not seek to investigate or police such suicidal tendencies(though this could be an aftermath of possibilities and use-cases, it is not the primary concern of this research), rather what this research mainly seeks to do, is only to understand a given piece of text to be able to predict if the person who wrote it has suicidal tendencies.

Furthermore, in the aftermaths (use-cases and possibilities) written briefly in the previous paragraph, it is fair to say that if the outcomes of this research were to be deployed in such manner, the person involved would have to give his/her/their consent.

Also, Twitter already has an extensive data privacy policy and is putting a very tight lid on how it's data is used for this research. Researchers now have to explain to Twitter in clear and concise manner how one is going to use their data. Consequently, any application of the NLP model that would be built would also have to follow the strict ethical guidelines and concerns.

One major limitation of this research is that its data is not varied enough. The research has it's focus and sights on just using User's Tweets as a source of data. This may make the model overfit on data only from Twitter. Depending on the angle you see this, this can be both a strength and a weakness. It is a strength in that the research has a given and specific problem, and hence stands a better chance of solving it; on the other hand, it's weakness could be that in other real-life scenarios or worlds outside User Tweets, the model may not perform as well. This is an assumption, and only till after the analysis stage when one performs a few experiments with the test data, would one know the answer definitely.

5 SUMMARY AND CONCLUSION

It is estimated that more than 700,000 people die due to suicide every year(who.int 2021). Its impact not only on the economy but in the primary unit of the family cannot be understated. This research seeks to help PCP better in identifying suicidal tendencies in text – especially in User Tweets since most people have an active Digital Phenotype and one can be able to be proactive in preventing suicide. But there are also concerns, one of which is do individuals have a right to end their life? This is especially relevant in the current climate where the US judiciary government is seeking to overturn an earlier decision that makes abortion legal in the USA(Josh Gerstein 2022); such a scenario can also play here ethically and morally. If I want to end my life, is it my choice? If yes, what would be the impact of this? The purpose of this research however is not euthanasia but rather to provide a model or more appropriately a tool for identifying suicidal tendencies. The philosophical impact and ramifications do seem intriguing and over the course of the research, it would be delicately threaded.

References

Bidargaddi, N., Musiat, P., Makinen, V.-P., Ermes, M., Schrader, G. & Licinio, J. (2017), 'Digital footprints: facilitating large-scale environmental psychiatric research in naturalistic settings through data from everyday technologies', *Molecular psychiatry* **22**(2), 164–169.

Boers, E., Afzali, M. H., Newton, N. & Conrod, P. (2019), 'Association of screen time and depression in adolescence', *JAMA pediatrics* **173**(9), 853–859.

Clayton, R. & Clayton, C. (2022), 'Uk screen use in 2022: A need for guidance'.

Coppersmith, G., Leary, R., Crutchley, P. & Fine, A. (2018), 'Natural language processing of social media as screening for suicide risk', *Biomedical informatics insights* **10**, 1178222618792860.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

Hall, A. (2018), 'Average briton gets six hours and 19 minutes of sleep a night, study finds'.

Horev, R. (2018), 'Bert explained: State of the art language model for nlp'.
URL: *https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270*

Josh Gerstein, A. W. (2022), 'Supreme court has voted to overturn abortion rights, draft opinion shows'.
URL: *https://www.politico.com/news/2022/05/02/supreme-court-abortion-draft-opinion-00029473*

Mikheev, A. (2003), Text segmentation, *in* 'The Oxford handbook of computational linguistics', Oxford Handbook of Computational Linguistics.

Pennington, J., Socher, R. & Manning, C. D. (2014), Glove: Global vectors for word representation, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1532–1543.

russelljohn.net (2008), 'A collection of suicide notes & letters'.
URL: *https://russelljohn.net/journal/2008/03/a-collection-of-suicide-notes/*

samaritan.org (2015), 'The samaritans radar twitter plug-in was closed permanently in march 2015'.
URL: *https://www.samaritans.org/about-samaritans/research-policy/internet-suicide/samaritans-radar/*

samaritan.org (2020), 'Latest suicide data'.
URL: *https://www.samaritans.org/about-samaritans/research-policy/suicide-facts-and-figures/latest-suicide-data/*

Scott, W. (2019), 'Tf-idf from scratch in python on a real-world dataset'.
URL: *https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089*

who.int (2021), 'Suicide'.
URL: *https://www.who.int/news-room/fact-sheets/detail/suicide*