STAY LATE AND CODE (SLAC 2020)

MENTAL HEALTH ANALYSIS

PROJECT REPORT ON

SENTIMENTAL ANALYSIS ON SUICIDAL CASES USING MACHINE LEARNING AND FLASK

Nakshatra Singh, Sri Suvetha C S,Sree lekha, Pradhikshen N

Submitted to SLAC 2020

8th November, 2020

Amrita School of Engineering,Bangalore

ABSTRACT--Worldwide, suicide rate is considered one of the most significant issue. With each passing year, the number of suicide is going up phenomenally and because of this reason, this research is carried out to predict the causes of suicide by using the machine learning algorithms and data mining techniques in order to identify the root causes behind the suicide so that the authorities can take advantage in order to prevent the suicide cases by creating awareness and by rectifying the predicted causes of suicides.According to a research, about 800,000 people commit suicide worldwide every year. Out of these, 135,000 (17%) are residents of India, a nation with 17.5% of world population.In this research, we have analyzed the pattern of suicide cases and predict the causes of future suicides by using machine learning algorithms and artificial intelligence. This paper studies the prediction of suicide by using machine learning method and techniques. Although ML has been a part of the computer science field for many decades, it has only recently been applied to clinical psychology.

KEYWORDS--Suicide,Machine learning

I.INTRODUCTION:

In this era of the internet and virtual world one could do anything with just a pc or a phone in hands . The Internet is like a data house of millions and millions of things where you can learn , create, explore,build, update and what not. It just turned our existence upside down with revolutionized communications and which is probably the most preferred medium of communication in everyday life. And today a click or two is enough to know anything about any corner of the world . Now let's count the number of online social platforms through which people could connect with each other.We know it's not a job of counting on fingers because there are countless platforms like twitter, facebook , linkedin , Instagram , whatsapp , etc and each one of these are renovating , modernizing in their own style.So looking at the positive side of this internet world technologies and social media platforms . The beneficiary list is pretty much endless. So let's focus on social media in everyday life. These are web based online tools that enable people to discover , learn , share ideas and interact with new people and organizations. It literally took communication to the next level. By making our lives much much easier.like consider just how easy it is to see what is happening on the other side of the world, through the accounts of real people rather than filtered news channels.Many of us wake up , turn off alarms , roll over and check what is trending on social media. Twitter in particular is a go to source for trending. It’s a platform that offers people to voice their opinions and it’s a go to go place to watch protests unfold and follow ,comment and a lot more. Here millions and millions of people are up with their own opinions , judgements , debates and all and it has a huge potential to discover and analyse infinite social media data for business driven applications. Social networks offer a considerable amount of content generated by the user, it is important content for analysis and offer more services adapted to the needs of users. In recent years, the majority of developments in the field of information and opinion exchange have launched the research work for the analysis of feelings expressed on these social networks so this made twitter to be more considerate for case studies like sentimental analysis. So here in this project we are working on suicidal predictions using sentiment analysis.Because suicide is simply one of the leading causes of death in the world . the pronunciation of suicide word should not be taken with simplicity and lightly. It can be the last cry of someone's help,

and yet if the signs and clues are recognized at the beginning, lives could be saved. Suicide is preventable, it is an act of those who have not been able to accomplish others and the prevention of suicide should be the responsibility of everyone. The purpose of this paper is to propose a method of predicting suicidal ideas, to predict suicidal acts and ideas using data collected from social media.

II. TWITTER SENTIMENT ANALYSIS

A) Introduction to Problem

In today's world most of us are stressed and depressed because of using mobile phones and other gadgets. In 2018 it has been stated that around 2.65 billion people were using social media worldwide , number thought to increase to 3.1 billion by 2021. Often people who intend to harm themselves will make an explicit statement to their social media circle or text to specific individuals regarding killing themselves or shooting themselves.Every day massive amounts of data is generated by social media users which can be analyzed to predict the suicidal thoughts of the normal people.

B) Platform and Technologies

Twitter:

It is an online social media platform which is suitable for our use case due to a number of factors. Firstly, the amount of relevant data is much larger for twitter as compared to blogs or review websites. Secondly, response on twitter is general and prompt. Other social media giants like Facebook do not provide much data so using their public API was not considered. Finally, most twitter users voice their opinion about other people .

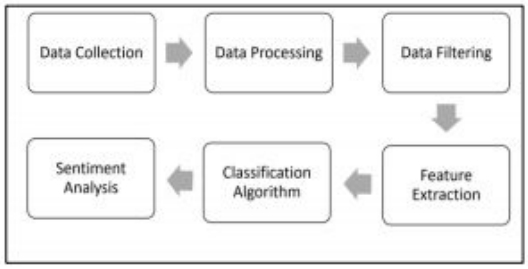
Google collab:

Google collab is a free jupyter notebook environment that runs entirely on the cloud. And one of the significant advantages of google collab is that it does not require any setup and more people can access it simultaneously. Google collab also supports many machine libraries which can be accessed without any installation in your computer and can be easily loaded to the notebook.Colab notebooks allow you to combine executable code along with images, HTML, LaTeX and more, and the notebook is autosaved to your google drive account so it can be accessed again.

Python3:

Python is a high level programming language used for general purpose programming it can be used for creating web application and connect to database systems and most absolutely rapid prototyping can be done using this language the reason we why we chosen this language is it is easy to read and maintain and it works in interactive mode and can it is extendable.

III. CASE STUDY:



1.DATA COLLECTION:

Datas for the sentimental analysis has be collected from kaggle,kaggle is a open source platform, subsidiary of Google LLC, which allows us to find and explore more datas related to machine learning and we have got a twitter sentiment analysis data from kaggle and contents for that data has been created from Reddit ,the data we extracted has been in the form of comma separated with text data and also the data has been binaryly classified into two categories that is suicidal and non suicidal.

2.DATA PROCESSING:

```
def clean_text(x):
  str(x).lower()
  re.sub(r'([a-z0-9+._-]+@[a-z0-9+._-]+)', '', x)
  re.sub(r'(http|https|ftp|ssh)://([\w_-]+(?:(?:\.[\w_-]+)+))([\w.,@?^=%&:/~+#-]*[\w@?^=%&/~+#-])?', '', x)
  re.sub(r'\brt\b', '', x).strip()
  re.sub(r'[^\w ]+', '', x)
  ' '.join(x.split())
  BeautifulSoup(x, 'lxml').get_text().strip()
```

Data processing is the process in which we are removing noise and preprocess tweets, the processing of data includes

- Converting the data to lowercase
- Eliminating all of the URL via regular expressions or replacing them with generic word URLs.
- Eliminating punctuation, special characters and additional white spaces at starting and ending of the tweets and multiple white spaces will be replaced by single whitespace.
  Ex:' this is awesome. ' will be changed as 'this is awesome'.

3.DATA FILTERING:

```
  BeautifulSoup(x, 'lxml').get_text().strip()
  ' '.join([t for t in x.split() if t not in stopwords])
  unicodedata.normalize('NFKD', x).encode('ascii', 'ignore').decode('utf-8', 'ignore')
  return x

df['tweet'] = df['tweet'].apply(lambda x: clean_text(x))
```

After data processing has been done the data will be still containing some raw information which will be not useful for our model so this data filtering we have removed some raw information like stop words. Stop Words refers to the common words in a language such as 'is, are , at, which ,on' so these kinds of words are removed from the data so it will be useful for us to train the model.

4. FEATURE EXTRACTION:

```
tfidf = TfidfVectorizer(max_features=30000,    # max 30000 dimensional vector space
                        ngram_range=(1, 3),    #unigram, bigram and trigram
                        analyzer='word')       #tokenization is done word by word
```

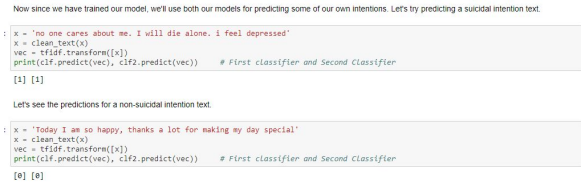
Here we are using *Tfidf Vectorizer*.In general Tfidf is a statistical measure that evaluates how relevant a particular information to the data. The TfidfVectorizer will tokenize documents, learn the vocabulary and inverse document frequency weightings, and allow you to encode new documents to a matrix of TF-IDF features.

Ex: If the tweet is 'Never flying your airline again!' the features of this tweet will be never flying,airline.

5.SENTIMENTAL ANALYSIS:

Sentimental analysis is the interpretation and classification of emotions with text data and text analysis technique.
And here we have used the SVM (Support Vector Machine) algorithm it tries to create a hyperplane which separates the data into two classes, here it separated as suicidal class and non suicidal class .Reason we have chosen SVM is it is good for analyzing sentiments for larger tweets and holds good for binary classification.

OUTPUT: After the algorithm calculator the which type of word is more prevalent in the text and if there are more positive words then the text is classified under the class of non suicidal or else it will be coming under the suicidal class.



V.FUTURE WORK:

For the future to truly understand and capture the broad range of emotions and express as written words, we need a more sophisticated multidimensional scale, and we are trying to implement this using different algorithms and improve the efficiency of the model and finally we are planning to implement this as a complete final product and give the users the better experience.

VI CONCLUSION:

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of corpus of texts very accurately because of the complexity in the language used , the grammatical issues and even more if we consider other languages . In this project we tried to show the basic way of classifying tweets into positive and negative. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of targets. Owing to numerous challenging research problems and a wide variety of practical applications, sentiment analysis has been a very active research area in several computer science fields. Here in this project we used this to predict suicides online.