ABSTRACT

Due to multiple reasons like the nuclear family, peer pressure for fake prestige, impatience attitude, and mental stress has become a common trait in every person. With advancements in technology like the internet and online presence, it has become a routine to be active online. Some sections of people vent out their emotions online as they have no support system in real life. It has been detected, as seen in some instances; those suicidal tendencies ranging from mild to extreme could be from a person's online profile activity. In our current work, we use a specific method that includes all critical criteria that could be exhibited by a suicidal person by using Natural Language Processing (NLP) techniques. NLP interprets written language, whereas Machine Learning makes predictions based on patterns learned from experience. These textual features are passed through a robust Machine Learning framework for detecting an abrupt change in input data. Our method predicts efficiently a genuine, mentally disturbed profile from a typical profile.

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE No
	ABSTRACT	I
	LIST OF FIGURES	II
	LIST OF TABLES	III
1	INTRODUCTION	1
2	LITERATURE SURVEY	5
3	METHODOLOGY	8
4	RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS	12
5	FUTURE WORK AND CONCLUSIONS	14
	REFERENCES	15
	APPENDIX	16
	A.SCREENSHOTS	16
	B. SOURCE CODE	17

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO
1	Prediction graph	05
2	Points and Line	06
3	Clustered data	07
4	Original Unclustered data	16
5	Select the Url	17
6	Detects without Suicidal content the url	17
7	Select the Url	17
8	Detects without Suicidal content the url	17

LIST OF TABLES

TABLE NO.	TABLE NAME	PAGE NO
1	Prediction Table	17

CHAPTER 1

INTRODUCTION

At the moment, the task of preventing suicide is relevant, because according to statistics of the World Health Organization every year more than 800 000 people commit suicide. According to the same data, the Russian Federation is among the five countries with the highest number of suicides per 100,000 inhabitants. Since the Internet is the easiest way to distribute suicidal content in the form of various web pages dedicated to suicide, a large number of organizations are trying to solve this problem, for example, the popular social network Facebook, which detects suspicious profiles of users prone to suicide, and posts containing suicidal content. In addition to social networks, federal services are trying to solve the problem, for example, in Russia Roskomnadzor in 2016 published recommendations for disseminating information about suicide cases in the media, which probably affects the results of search engines on this topic. In addition, since 2006 there has been a unified register that contains websites blocked in the Russian Federation. However, blocking does not happen immediately. Some people manage to visit dangerous web pages. Manual blocking of suicidal websites can hardly be called an effective measure in the fight against the spread of suicidal content. Since after several years of such blockages, the number of child suicides in Russia has risen sharply again. In addition to the «death groups» in social networks that platform developers are already actively fighting, one of the reasons may be that websites with suicidal content can create their own copies (mirrors). This article discusses the possibility of detecting such web pages by analyzing their content in real time using machine learning algorithms. Detection occurs on the client's side. In this way, with sufficient accuracy to identify dangerous websites visited by the user, it is possible to identify a person who is suicidal at an early stage.

CHAPTER 2

LITERATURE SURVEY

The SGD algorithm is good at identifying harmful websites, but it can make mistakes when evaluating which ones are safe, labelling them as risky.[1]The web mining algorithm will extract textual information from web pages and identify those associated with terrorism. A system whose main purpose is to create a website where people may inspect any webpage or website for any evidence of terrorist activity.[2]The persuading concept is to see if the feature's equivalent word appears in the mail. When a good classifier is employed to create the classification model, the experimental results of this method have high TPR and Precision values, and the false positive rate is regulated within an acceptable range.[3]Linguistic characteristics are critical for distinguishing across users' written styles. Sentiment analysis is most effective with content that has a subjective context, such as a suicide note. These characteristics can be derived explicitly from the user profile or inferred implicitly using various data mining tools and methodologies.[4]Using a document embedding, A decision tree model with gradient boosting predicts dangerous categories of gathered web pages. [5] The ROC curve was used to comprehend a performance measurement for a classification task at various thresholds. The false-positive percentage should be kept as low as possible, whereas the true positive rate should be maximized. [6]Hierarchical spatial scaling's analytic bias improves the model's ability to handle detection problems in documents of possibly changing sizes. A feature extractor is a programme that parses a neural network model that distinguishes a series of tokens from HTML pages judgments, in this approach.[7]Cross-channel scripting defence methods follow a website's whole path, consisting of sustained storage systems If the soiled information is not cleansed, an alert is generated. The adversary can use this threat to insert inappropriate material into the user's embedded system. causing web applications to malfunction and information to be leaked.[8]Clustering methods lustring (e.g., k-means, DBSCAN) To detect malicious domains, unsupervised separation instances evaluate data and derive necessary details from it.[9]The Dirichlet latent allocation topic model proposed a pattern that determines that tweets about crime are more likely to be positive. However, A stash of tweets is available; however, a series of old tweets is either impossible or prohibitively expensive.

CHAPTER 3

METHODOLOGY

We propose a system with the primary goal of creating a website where users can search for any trace of terrorist activity on any web page or website. To accomplish this, our website will allow users to enter the URL of the web page they wish to scan. After entering the URL, our system will count the words on the entire web page and compare them to the words already in our database. Each word that we store in our database will be assigned a score. Our system will retrieve the scores for each word on the user's web page from our database, and then it will compute the overall rank of the website. This rank will determine whether the user's website contains any traces of terrorism. Using web mining and data mining, our system will detect patterns, keywords, and relevant information in unstructured text on a webpage. To retrieve textual content, our system will employ a web mining algorithm from web pages to recognize those that are relevant to the case. Web mining and data mining are sometimes used in tandem to achieve the best results. The application was built using Python and machine learning algorithms. The prospect of discovering This kind of web page can be recognized via using machine learning algorithms to scrutinize its content in real-time as discussed in this article. The client is the one who notices the issue. In this approach, suicidal people can be recognized early enough to recognize unsafe websites that the user visits frequently. A machine learning model is nothing more than a piece of code that has been trained with data by an engineer or data scientist. So, if you feed the model garbage, you'll receive garbage back, i.e. the trained model will forecast false or incorrect.

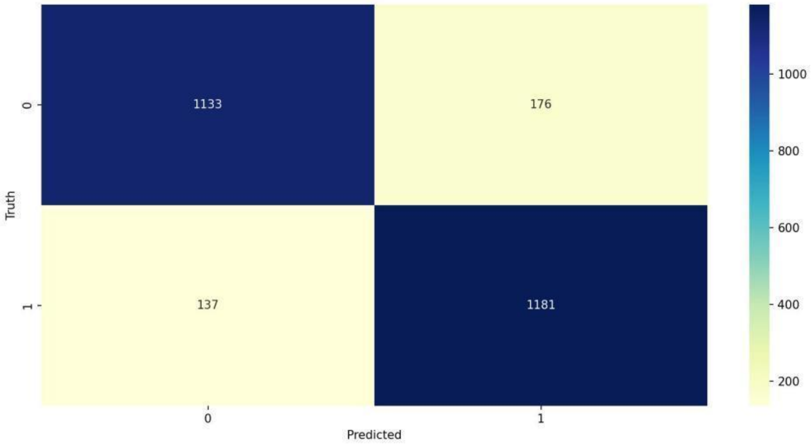

Fig1.Prediction graph

Beautiful Soup is a parsing package that performs a fantastic job at retrieving contents from URLs and allowing you to parse specific parts of them with ease. Web Scraping is also known as web data extraction, is a method of obtaining information from websites. The majority of this information is in the form of unstructured HTML that is converted to structured data in a spreadsheet or database before being used in various applications. It extracts data in a more comprehensible and hierarchical manner by generating a parse tree from the page source code. It's a complete web- scraping or crawling framework

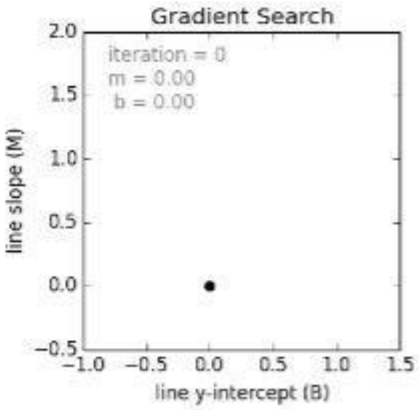

Fig2. Points and Line

Tkinter is a Python package that is frequently used to create graphical user interfaces. Tkinter makes creating a GUI a breeze, and the process is even faster. Tkinter provides a number of widgets that can be used to create a graphical user interface. Data collection enables us to keep a log of past details in order to use data analysis to uncover sequences. You can produce statistical models using machine learning algorithms that look for patterns and estimate future changes depending on those patterns. Good data collection processes are critical for generating high-performing models since predictive models are only as good as the data they're built on. The data must be devoid of errors (garbage in, garbage out) and contain information that is relevant to the work at hand. A debt default model, for example, would not profit from tiger population sizes but would benefit over time from gas costs. Much different statistical analysis and data visualization approaches are used to investigate data in order to determine which data cleaning activities should be performed. We created a pre-processing function that helped lowercase the text, remove punctuation, and lemmatize related terms down to a single base word. After the pre-processing, the texts are evaluated using a machine learning algorithm that has been pre-trained to categorise themas possibly lethal or safe. The portion of the maximum of terms calculated by the algorithm is then computed. A pre-prepared training set was used for training and precision testing, which was gathered and classified based on certain texts into training and test data. The training dataset, that includes roughly 700.000 texts, and was used to train the models stored in nearly 2.000 texts, was used to evaluate the accuracy of the algorithms

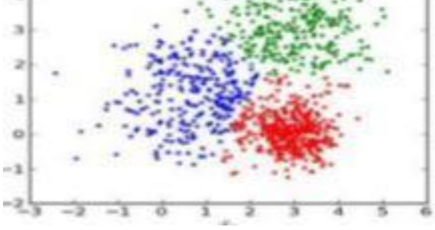

Fig3. Clustered data

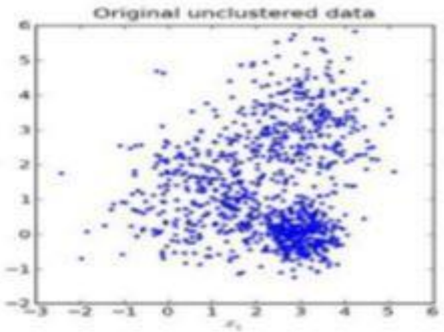

Fig4 .Original Unclustered data

To further comprehend our data, we employed a Count Vectorizer to examine the most frequently used words in each subedit (The words used were similar with some subtle differences). The changed text is then applied to various algorithms for training. Following that, a group on the web was constructed to test the algorithms under realistic conditions. Algorithms that are used are In its simplest form, logistic regression is a statistical method that applies a logistic function methodology is established to a binary dependent variable.There are numerous intricate augmentations In regression analysis, logistic regression (or logistic regression) is used to quantify a logistic model (a sort of binary regression). In statistical software, it is used to analyze possible outcomes and comprehend the correlation between the dependent and independent variables by using a logistic regression equation. This form of assessment can guide you in foretelling the likelihood of activity occurring or an intention being made. Beautiful Soup is a Python library for data extraction that retrieves data from HTML and XML files. It produces a parse tree from the page code base, which can be used to extract information in a more consolidated and readable manner. t is a complete framework for web-scraping or crawling. BeautifulSoup is a parsing library that also does a job of fetching contents from URLs and allowing you to easily parse specific parts of them. It only retrieves the contents of the URL you provide and then exits. Tkinter is a Python library that is commonly used to create graphical user interfaces (GUIs). It is very simple to create a GUI with Tkinter, and the process is even faster. Tkinter includes a number of widgets that can be used when creating a graphical user interface. These include buttons, radio buttons, checkboxes, etc The constant data process is performed in machine