Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Predicting Harmful Web Pages Based on Suicide-related Textual Analysis using Machine Learning Algorithms

A.Yovan Felix[1*], M Dharshini Nithila[2], Vinisha R[3], J. Jabez[4], P. Rajasekar[5]
[1]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
yovaanfelix@gmail.com
[2]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
datchu.guna@gmail.com
[3]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
vinizarsyn@gmail.com
[4]Department of Information Technology,
Sathyabama Institute of Science and Technology, Chennai, India
jabezme@gmail.com
[5]Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology Chennai, India
rajasekar.dr.25@gmail.com

Article Info
Page Number: 468 - 477
Publication Issue:
Vol 71 No. 3s2 (2022)

Article History
Article Received*: 28 April 2022*
Revised*: 15 May 2022*
Accepted*: 20 June 2022*
Publication*: 21 July 2022*

Abstract
And detecting suicidal people remains a difficult task. As the usage of social media has grown, we've seen people openly discuss their suicide plans or attempts on these platforms. Suicide prevention is addressed in this article by identifying suicidal profiles on social networks. First, we examine online profiles and extract a variety of information, such as account features connected to the profile and features relevant to the social media data. Second, we present our technique for detecting suicidal profiles using Twitter data, which is based on machine learning algorithms. Then, as a profile data set, we employ a data set of people who have previously committed suicide. The efficiency of our technique in terms of memory and precision in detecting suicidal characteristics is supported by experimental data. Finally, we demonstrate the detection of suicidal characteristics using a Java-based prototype of our work.

Keywords: - Machine learning, Logistic regression algorithm

I. INTRODUCTION

Presently, the mission of preventing suicide is important since, according to WHO estimates, over 800,000 individuals commit suicide each year. Russia is one of the five countries with the highest suicide rate for every 100,000 people, according to the records. Because the Internet is the most convenient means of disseminating suicidal information in the form of various suicide-related web pages, many organizations are attempting to address the issue, such as Facebook, which detects

suspicious profiles and posts containing suicidal data. In Russia, for example, Roskomnadzor established instructions for spreading there were a lot of data about occurrences in the networks during 2016, which is probably affecting search queries on this topic. Likewise, in 2006, the Russian Federation maintained a consolidated register of websites that are restricted. Blocking, on the other hand, is not instantaneous. Some individuals can use harmful websites. Manually censoring suicidal websites is rarely a viable strategy for combating the spread of suicidal content. The rate of juvenile Suicides in Russia is on the rise. Substantially again following several years of such blockades. Yet another reason might be that online sites with suicide stuff can create regular replicas, in addition to death groups in social networks, where platform developers are already openly supporting combating (mirrors). This article explores how machine learning techniques may be used to such web pages can still be detected by comparing their content in real-time. Detection occurs at the client's end. Suicidal users can be recognized early by using this technique with adequate accuracy in identifying risky websites frequented by the user.

II. LITERATURE SURVEY

The SGD algorithm is good at identifying harmful websites, but it can make mistakes when evaluating which ones are safe, labeling 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 program that parses a neural network model that distinguishes a series of tokens from HTML page judgments, in this approach.[7]Cross-channel scripting defense 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 clustering (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.

III. Related Works

Researchers used specialized search engines and found several active suicide forums. [10]Using ordinary least squares, a regression-based strategy was used to predict three types of crime using Poisson regression and negative binomial regression models. However, a Poisson distribution cannot be used to match the data. [11]An algorithm that spreads and gathers harmful Web pages selectively based on Web server aspects relating to the Target Website's URL. When the process is unable to discover a URL with an EML value greater than the threshold, the suggested technique, unlike existing algorithms, can terminate crawling properly. [12]N-gram determines the number of times each byte code appears in a string. The use of byte codes is checked for entropy. If there is an extremely long string, the word size is checked. A new technique for detecting obfuscated strings in malicious websites has been developed. [13]The J4.8 machine learning system detects suspicious websites using simply the domain extensions. The learned knowledge must be updated regularly as attackers' strategies change.[14]Although k-means clustering identified and solved the absence of negative samples in various datasets, The framework is beneficial to the estimation of crime hubs as well as other analysis work domains. [15]

IV. Research Motivation

With the population increasing, it's much more necessary than ever to identify and treat depression in older individuals. Machine learning techniques are improving rapidly in assessing, monitoring, and predicting depression and other comorbidities in both young and old people. It concentrates on analyzing the acoustic and linguistic characteristics of human language generated from speech and text, and they can be combined with machine learning methods to categorize depression and its severity. Greater biological validity, low subjectivity, low cost of frequent assessments, and speedier task administration compared to routine assessments are all advantages of adopting these approaches to comprehend depression symptoms through speech.

V. Proposed system

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.

VI. METHODOLOGY

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.

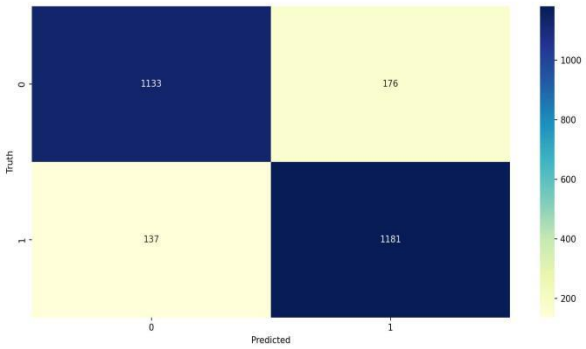

Fig1. Prediction graph

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 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

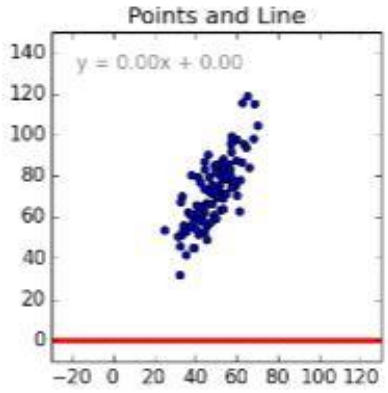

Fig2. Points and Line

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 several 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.

Much different statistical analysis and data visualization approaches are used to investigate data 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 categories them as 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, which 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

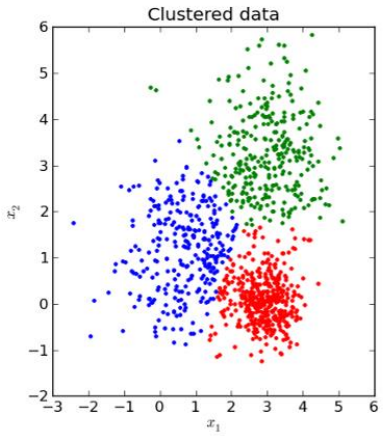

Fig3. Clustered data

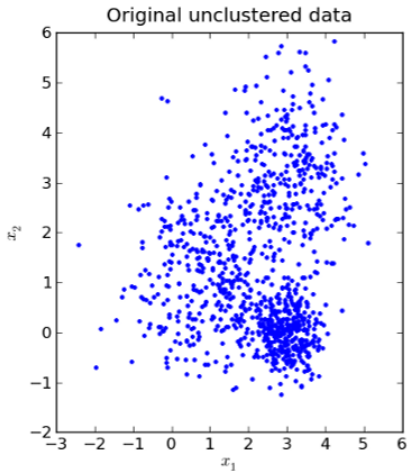

Fig4 .Original Unclustered data

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).Statistical software 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 an 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. Beautiful Soup 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 several 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 learning, and Python's libraries allow you to access, handle, and modify information. These are the most extensively used libraries for ML and AI: Scikit-learn is used to handle simple machine learning algorithms such as clustering, linear and logistic regressions, regression, classification, and others.

VII. PERFORMANCE ANALYSIS

ALGORITHM	FUNCTION	LIMITATION	PERFORMANCE ANALYSIS
Rotation Forest algorithm and	Maximum Probability voting classification decision method	This approach works best with users who mention their age.	Accuracy is high
Support Vector Machines and Logistic Regression	Cross-validation method	The efficacy of Twitter for actual suicide ideation is unidentified, and the study results do not directly describe intervention targets.	Accuracy score of 76%
Support Vector Machines (SVMs) with Term Frequency weighted by Inverse Document Frequency	Naive Bayes algorithm	There is no connection between the timing of suicidal tweets and real suicide.	Achieved a performance of 0.657
Transfer learning	Discrete distance decay function	DNN-based method for the prediction of crime occurrences cannot be applied when sufficient data is unavailable.	DNN model is a more accurate

VIII. System Architecture

Firstly the data is gathered depending upon the work we want to do and it is sent to the Preprocessing unit where cleaning of the raw data takes place like removal of stop words, punctuation marks, special characters and stemming is done. Then it moves to the Data split unit where the corpus is divided into Test and Training sets, where it learns how to process the information. It is then sent to the Feature Extraction part.

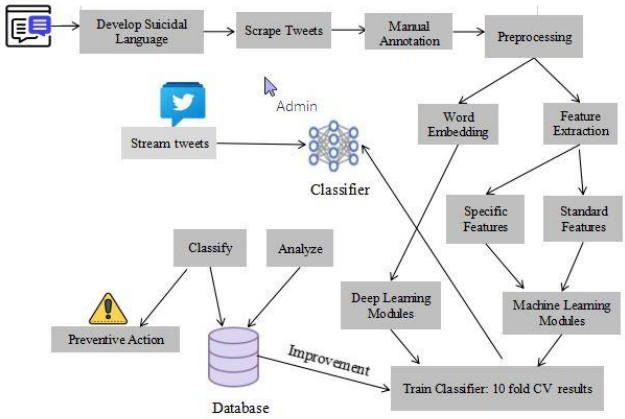

Fig5. Architecture diagram

There it is extracted into TF-IDF and Word2vec. This data is sent to the Classifier/Learning model and the Prediction part. In the Classifier/Learning model, a python library is used for web scraping purposes to pull the data out of HTML and XML files called Beautiful Soup. Moving to the Prediction part the data is Tested and Trained and sent to the last part which is the Evaluation Parameter where Accuracy, Precision, Recall and F-measure are done.

IX. Result

In this concept design study, we suggested a technique for creating a classification model for precedent suicidal ideation among adolescents using natural language processing and machine learning of clinical narratives from health record data before admission. This is the first time that unstructured data has been used to up skill a machine-learning classification system in a teenage inpatient population using NLP as far as we know. This first success signal based on a limited sample size needs to be confirmed in bigger datasets, The method demonstrates how EHR notes for adolescent suicide attempts can be enhanced with clinically relevant information to identify children with a history of suicide risk, This can aid in patient's health organizing during a highly vulnerable period for this high-risk demographic.

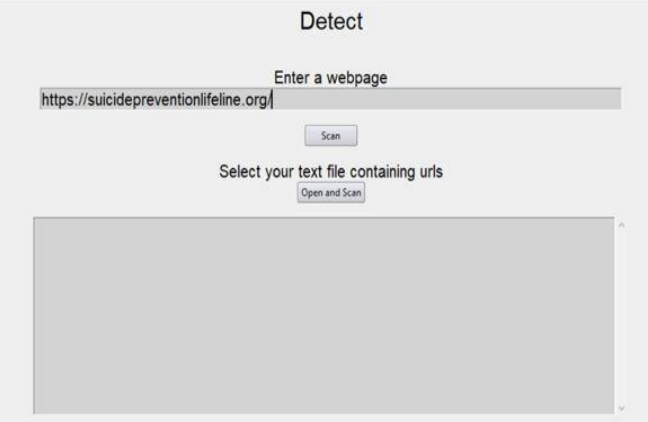

Fig 6. Select the Url

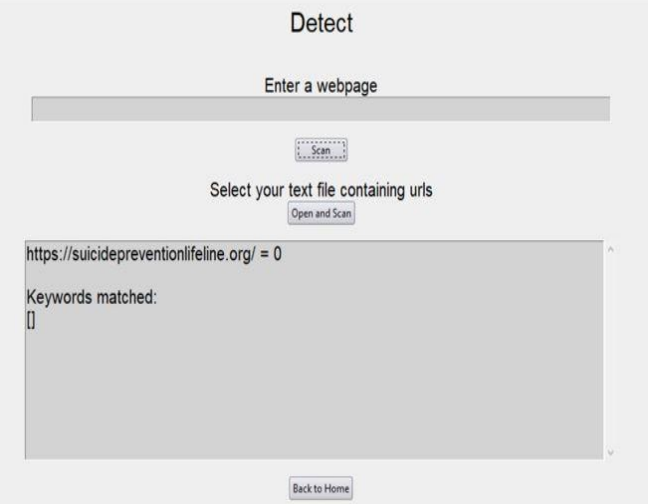

Fig 7. Detects without Suicidal content the Url

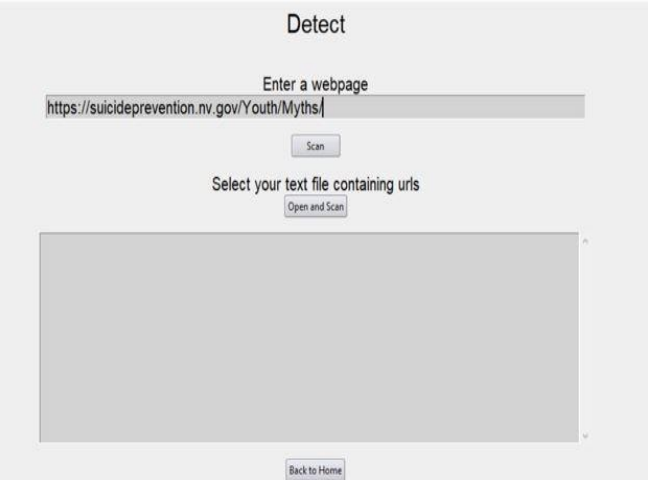

Fig 8. Select the Url

Fig 9. Detects Suicidal content the Url

X. CONCLUSION

This paper created a cyber-bullying detection model using TF-IDF and Word2Vec feature extraction. Several machine-learning-based text classification approaches were also investigated. The tests were carried out with the help of a global Twitter dataset. According to the analytical outcomes, LR has the highest accuracy rate and F1 score in our dataset. With 90.57 percent classification accuracy and 0.9280 F1 scores, respectively. Meanwhile, the performance of the LR, SGD and LGBM classifiers differs slightly, with SGD achieving 90.6 percent accuracy but a lower F1 score than LR.

The LGBM classifier, on the other hand, had an accuracy of 90.55 percent and an F1 score of 0.9271. This indicates that LR outperforms other classifiers. Furthermore, it was discovered during the trials that LR performs better as data size grows and when compared to the other classifiers used in this study, has the shortest prediction time. As a result, SGD performs nearly as well as LR, although the error is not as small as it is in LR. Feature extraction is an important part of machine learning for improving detection precision we did not examine many extracting features methods in this study. one more area for improvement to achieve a more accurate rate of both the LR and SGD classifiers is to integrate and compare different feature extractions. Another constraint on which we are working is the creation of a real-time cyber bully prediction tool that will aid in the detection and prevention of cyberbullies in real-time. Another line of investigation is cyberbully detection in many languages, with a focus on Arabic.

XI. REFERENCES

[1] Lyovkin, Maxim, Aleksey A. Frolov, and Egor Perminov. "Detection of Dangerous Web Pages Based on the Analysis of Suicidal Content Using Machine Learning Algorithms."*2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. IEEE, 2021.

[2] Goradia, Rinkle, et al. "Web Mining to Detect Online Spread of Terrorism."*International Journal of Research & Technology (IJERT)* 9.7 (2020): 645-648.

[3] Li, Xue, Dongmei Zhang, and Bin Wu. "Detection method of phishing email based on persuasion principle."*2020, IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*.Vol. 1. IEEE, 2020.

[4] Mbarek, Atika, et al. "Suicidal Profiles Detection in Twitter."*WEBSITE*. 2019.

[5] Kawaguchi, Yuki, and Seiichi Ozawa. "Exploring and identifying malicious sites in dark web using machine learning."*International Conference on Neural Information Processing*. Springer, Cham, 2019.

[6] NARYNOV, S., MUKHTARKHANULY, D., KERIMOV, I. and OMAROV, B., 2019.“Comparative analysis of supervised and unsupervised learning algorithms for online user content suicidal ideation detection.”, *Journal of Theoretical and Applied Information Technology*, *97*(22), pp.3304-3317.

[7] Saxe J, Harang R, Wild C, Sanders H. “A deep learning approach to fast, format-agnostic detection of malicious web content”. In2018 IEEE Security and Privacy Workshops (SPW) 2018 May 24 (pp. 8-14). IEEE.

[8] Madhusudhan, R. "Cross Channel Scripting (XCS) Attacks in Web Applications: Detection and Mitigation Approaches."*2018 2nd Cyber Security in Networking Conference (CSNet)*. IEEE, 2018.

[9] Buber, E., Demir, Ö., & Sahingoz, O. K. (2017, September). Feature selections for the machine learning-based detection of phishing websites. In *2017, international artificial intelligence and data processing symposium (IDAP)* (pp. 1-5). IEEE.

[10] Gerber, Matthew S. "Predicting crime using Twitter and kernel density estimation."*Decision Support Systems* 61 (2014): 115-125.

[11] Shingleton, Jarrod S. *Crime trend prediction using regression models for salinas, California*. NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 2012.

[12] Hattori, G., Matsumoto, K., Ono, C., & Takishima, Y. (2010, October). Identification of malicious web pages for crawling based on network-related attributes of the webserver. In *2010 4th International Universal Communication Symposium* (pp. 355-361). IEEE.

[13] Choi, Y., Kim, T., Choi, S. and Lee, C., 2009, December. Automatic detection for javascript obfuscation attacks in web pages through string pattern analysis. In *International Conference on Future Generation Information Technology* (pp. 160-172). Springer, Berlin, Heidelberg.

[14] Seifert, Christian, Ian Welch, Peter Komisarczuk, Chiraag Uday Aval, and Barbara Endicott-Popovsky. "Identification of malicious web pages through analysis of underlying DNS and web server relationships.", In *2008 33rd IEEE Conference on Local Computer Networks (LCN)*, pp. 935-941. IEEE, 2008.

[15] Felix, A. Yovan, and T. Sasipraba. "Decision support system for flood risk assessment and public sector performance management of emergency scenarios." International Journal of Public Sector Performance Management 8, no. 3 (2021): 219-229