International journal of applied exercise physiology
ISSN: 2322-3537
Vol-14 Issue-01 June 2025

Predicting Suicidal Thoughts by Analyzing Twitter Data using Machine Learning Techniques

[1]Pushpa Soumya, [2]M. Lohitha

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract—

Thanks to the proliferation of online social networks and microblogging services like Twitter on the World Wide Web, information may now reach vast audiences in a matter of seconds. An alarming example of this kind of online collaborative contagion is the spread of harmful ideas on social media platforms such as Twitter. Using a point-order analysis of tweets and phrases related to suicide, this inquiry discusses the implications and results of several machine classifiers. More distressing data, such as suicidal ideation, more suicide-related themes, detailed facts about suicide, allegiance, campaign, and support, may be distinguished by the classifier. The utilization of Twitter's emotional, linguistic, psychological, and structural features in a basic classifier allows for the linking and identification of suicide references. Machine learning methods such as clustering, bracketing, association rules and natural language processing (NLP) are used in this operation. To help direct future studies in this area, this investigation looks at the challenges and limitations already present.

Index Terms—Machine Learning, Natural Language Processing (NLP), Data Mining, Classification, Python, Tokenization, Lemmatization, Text Analysis

INTRODUCTION

Anxiety and depression are major public health concerns in the modern world. Developed countries are hit the worst by them. People with severely dysfunctional mental disease are finding an increasing audience in these fast-growing countries. Help is available for those struggling with suicidal thoughts and behavior. Cyber bullying and online stalking are problems that might arise because some internet communications include a lot of inappropriate content. Offensive information like this is often seen on social networking platforms. Brutality, gossip, and mental anguish are the outcomes. Suicide and cyber bullying may have a connection, according to the available information. It is possible for victims to feel depressed and frustrated after experiencing several negative messages and events. Even worse, individuals do commit suicide from time to time. The reasons for suicide might vary. Suicidal ideation is common among those who do not suffer from depression, but those who do are more likely to take their own lives. Suicide has three main causes, according to the American Suicide Prevention Foundation (AFSP): environmental, historical, and health-related. Mental illness and substance abuse are associated with an increased risk of suicide, according to research by Ferrari et al. Individual and personal psychological risk variations, cognitive variables, social influences, and unpleasant life experiences were the four main points of O'Connor and Nock's comprehensive analysis of the psychology of suicide. A person's suicidal thoughts or sentiments may be determined by suicidal ideation detection (SID) using either tabular data or written material that the person has contributed. With the rise of social media and the ability to remain anonymous online, more and more individuals are interacting with others in this way. Suicidal ideation, emotional distress, and physical agony are all too common among those who express themselves online. It was just a matter of time before internet platforms started keeping tabs on people's suicidal ideas and mining social media for data that may aid in suicide prevention. Online groups that promote copycat suicide and self-harm are only one of the many unique social phenomena that are growing in popularity every day. Suicide prevention has, similarly, received a great deal of attention as of late. Indeed, the ease with which one may acquire depressing emotional notions and attitudes is one among the most striking features of social networks. This is why many academics use social media to learn more about suicide. For instance, Millions of Twitter users keep expressing themselves via brief postings called tweets that include semantic expressions like emoticons, hash tags, special characters, etc. This means text mining can tap into a plethora of Twitter data.

The majority of those who consider suicide use some kind of social media to broadcast their plans to others. You could hear things like "I'm so tired," "I hate my life," "I have lived long enough," or "I want to end my life" (among others). The fastest way to stop suicide attempts is to find these warning signs and other hidden signals that are behind their posting content. Then you can respond and stop their deaths. The profile photo and tweet stream of a Twitter user may frequently be used to identify them. Some of the characteristics that define a person in a profile include their name, age, place of residence, and birthdates. Despite the wealth of information available in tweets for facial recognition, many important characteristics about a user's public profile that may aid in suicide detection go unnoticed. Efforts to tackle the problem of suicide profile identification also make use of user-shared data, sometimes known as account characteristics, and tweets in the process. To start, it is not easy to glean further information about specific people from their tweets. When it comes to user-posted tweets, the most challenging semantic features are the ones that aren't easy to directly extract, including n-grams, stylometry, writing style, emotions, emoticons, hash tags, etc. Because a user's tweets may be a reflection of their personality and habits, we analyze and extract as many semantic features as we can from the tweets. Proper analysis of publicly available tweets yields grammatical, emotional, stylometric, and other semantic aspects. With these characteristics, we can tell different people's writing styles apart, which helps us identify if they are suicidal or not. The extraction process makes use of a wide range of data mining techniques and instruments. Several categorization approaches are used and a supervised machine learning model is used to learn the features of each user's tweet so that it may be labelled as suicide or not. The goal of this test is to identify trends among Twitter users who have expressed suicidal thoughts or who have already taken their own lives.

EASE OF USE

A. Current Layout By examining the Twitter accounts of 135 research participants, Braithwaite et al. validated an ML method to identify sad persons with 92 percent accuracy using the Linguistic Inquiry and Word Count (LIWC) technique. Messages heavily mentioning suicide had distinct language patterns, as shown in 2017 by O'Dea et al. using LIWC analysis on Twitter data. Furthermore, a Convolutional neural network that can identify suicide-related tweets was trained using ML by Du et al. in 2018. Burnap et al. developed a machine learning classifier that can, with an estimated accuracy of 68-73 percent, identify Twitter users with SI. By recognizing tweets that mention suicide or include the word "suicide," these strategies open the door to the possibility of rapid interventions in those who are at risk of trying suicide. An accuracy of around 85 percent may be achieved by using a binary classification of those who self-harm; one's message has to avoid mentioning suicide in order to be categorized. Posts are categorized as suicidal or non-homicidal using a credible classifier and the suggested approach exhibits refined accuracy, as described before. Using advanced tools like the NLTK library in Python with an interactive integrated development environment (IDE), such as Jupiter Notebook, has made tweet analysis, labeling, and prediction easier, faster, more dependable, and more accurate.

LITERATURE SURVEY

The primary objective of this research is to detect suicidal ideation by mining the Twitter database for a large number of tweets, analyzing each one word by word, and assigning a suicide risk rating to each one. This was accomplished by compiling a small number of publications written by subject-matter experts and by consulting a large number of articles. Recognizing possible suicide tweets or not is partly related to the use of sentiment evaluation by machine learning and natural language processing in text analysis, which is

International journal of applied exercise physiology
ISSN: 2322-3537
Vol-14 Issue-01 June 2025

the focus of this study review. Md. Taufiqul Haque Khan Tusar and Md. Touhidul Islam presented a study on sentiment evaluation using natural language processing and several machine learning approaches using Twitter data from US airlines. Determining whether textual data is positive, negative or neutral is the primary objective of sentiment analysis. Using sentiment analysis tools, decision-makers may track how the public or consumers view different people, objects, activities, technology, and services over time. Businesses, political groups, and nonprofits may all benefit greatly from sentiment research, which can help them improve their goods and services more quickly. Thanks to sentiment analysis, it's much easier to grasp the public's opinion in a flash. Databases called datasets contain the majority of the data used for sentiment analysis, the majority of which originates from social networking websites. Sentiment analysis may be challenging with imbalanced, massive, multi-classed, etc. datasets. Twitter US Airline Sentiment, a large, unstable, multi-class, real-world dataset, was used in their study. The data had already been vectorized and pre-processed using natural language processing techniques. Next, the polarity of the text was categorized using machine learning techniques for categorization. To determine the best move, we contrasted applicable NLP techniques with Machine Learning algorithms. An independent review was conducted by Arwa Alshamsi of The British University in Dubai's Faculty of Engineering and IT, Reem Bayari of the University of Sharjah's Research Institute of Sciences and Engineering, and Said Salloum of the University of Salford's School of Science, Engineering and Environment in Manchester, UK. Businesses may learn a lot about their customers' opinions about their own and competing brands from the data provided by social media sites like Instagram, Twitter, and Facebook. In addition, decision-makers interested in improving the services given are attracted to these helpful statistics. This research paper reviewed a number of studies that examined Twitter's data classification and analysis for different purposes in order to learn more about the methods and procedures utilized for text categorization. The authors of this study looked for open-source datasets before testing several text categorization methods and machine learning methodologies. The authors used a number of classifiers to sort texts from two different dataset versions. From a scientific perspective, the two most common ways to detect suicidal thoughts are via machine learning and questionnaires. Stephanie et al. found that there is a wide variety of effective scale-based models and assessment questionnaires for predicting STB. But some do write "suicidal" statements on platforms like Twitter, which provides more real information, due to the increasing use of social media today. Natural language processes and machine learning have seen widespread usage of innovative methodologies, which has also made it possible to extract semantic information from voice and text. As a result of this development, new linguistic avenues for STB prediction have opened up. Using data collected from online social media platforms, we can identify suicide posts and help those in need. A machine learning-based approach for detecting suicidal thoughts, applied to a publicly accessible dataset of micro blogging posts.

ARCHITECTURE AND METHODOLOGY

Here are the many stages of the methods: FIRST STEP: Gathering Data for Machine Learning Classifier Evaluation and Training. 2) Preparing the dataset for further processing by cleaning it up. Making use of natural language processing to convert text to vector format. 4) Separating the dataset into two parts: training and testing. By comparing the experimental data to the training data, the ML Classifier may learn to predict the polarity of the data. The analysis and classification processing flow is shown in Fig. 1. A. Gathering Information the data sets used for the analysis, extraction process, and classification of tweets were mostly sourced from the Kaggle machine learning repository, an open data source for data mining and predictive analytics. The data needed for this project was gathered and analyzed using a pre-existing data collection on the Kaggle website. A CSV file containing the user's Twitter handle and all of their published tweets

International Journal of applied exercise physiology
ISSN: 2322-3537
Vol-14 Issue-01 June 2025

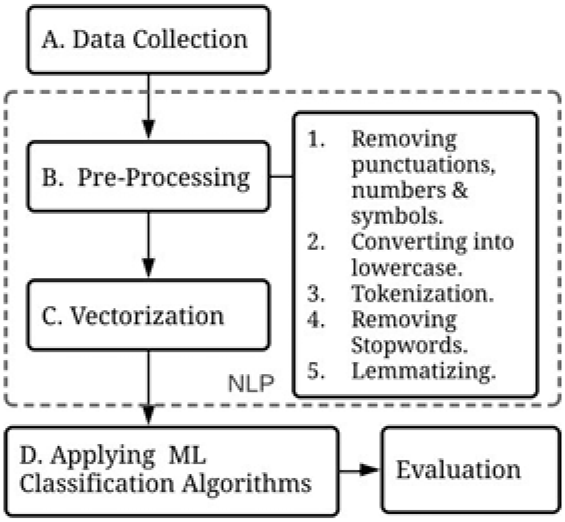

Fig. 1. Flow process of analysis and classification of tweets on the website was utilized as the data source for the acquisition.

Data Purification Checking the contents of the CSV file for missing values, duplicated information, or conflicting data is critical since the data source could provide inconsistent data. There must be a thorough evaluation and confirmation of all variables using filters, with the elimination of null values, to ensure the highest degree of accuracy. C. Initial Steps Tweets may include more than one character (! @, etc.), numbers, punctuation, and stop words. When we talk about words like "he," "she," "the," "is," and "that," we're referring about stop words. Stop word elimination is done using the NLTK library. Rich text pre-processing is made easy using the Natural Language Toolkit, sometimes known as NLTK. In NLTK, we have a collection of stop words in sixteen different languages. Importing the class "stop words" from this extensive library and specifying the language or words you desire to exclude will remove it. The use of noisy data in sentiment analysis is nothing new. To get the data suitable for processing, we remove any punctuation, numbers, and symbols and convert all the characters to lowercase. You may change a string of uppercase letters to a lowercase one using a built-in Python function.

Table I
PRE-PROCESSING OF NOISY DATA

Tweet 1	#Delicious #Beef #Cheese #Burger @McDonald Testing CheeseBurger and Hamburger
After Pre-processing	[delicious, beef, cheese, burger, mcdonald, taste, cheeseburger, hamburger]
Tweet 2	#Late Service @McDonald Delicious Hamburger but slow service
After Pre-processing	[late, service, mcdonald, delicious, hamburger, slow]
Vocabulary	[delicious, beef, cheese, burger, mcdonald, taste, cheeseburger, hamburger, late, service, slow]

Minus one. This also applies to strings that include both uppercase and lowercase letters. Using the "lower ()" method, strings are transformed into lowercase. Decomposing a continuous stream of words into individual words or tokens is called tokenization. Tokenizing text is made easy by using whitespace as a "delimiter" of words in a string. Tokenization was an important part of this study's methodology, and the split function in Python was useful for doing this. The tweet was then tokenized, and the resulting list of tokens was subsequently cleaned of stop words and stripped of any further meaning. A lemmatized form has been transformed from a basic one. A list named tweets was then created to contain the basic forms of each cleaned and prepared tweet. You may see the results of the preprocessing in Table I. D. Automated Language Recognition (ANN) In computer science, the subfield of AI known as "natural language processing" (NLP) aims to improve computers' comprehension of written and spoken language to a level comparable to humans. A combination of statistical significance, neural networks, deep learning models, and rule-based models of human language constitute natural language processing (NLP). Thanks to these advancements, robots can now "understand" what is being said or written, including the speaker's or writer's intentions and emotions, and can even transpose human speech into text or audio data. Many software systems rely on natural language processing (NLP) to do tasks such as translating text across languages, responding to spoken requests, and rapidly summarizing large amounts of text, sometimes even in real-time. To aid the computer in understanding the written and spoken data it is consuming, several NLP procedures deconstruct voice and text data. Examples of what they stand for include natural language generation, sentiment analysis, part-of-speech tagging, co-reference resolution, named entity

International Journal of applied exercise physiology
ISSN: 2322-3537
Vol-14 Issue-01 June 2025

recognition, word meaning differentiation, and word segmentation.

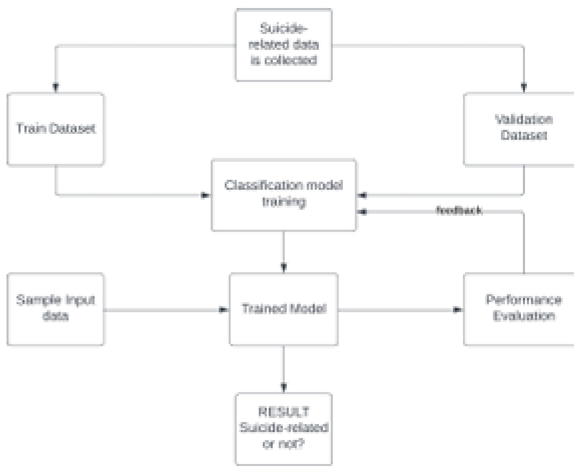

Fig. 2. The system architecture of the proposed design

Feature Extraction

A straightforward step in Natural Language Processing that helps us understand the context better is Feature Extraction. Feature extraction is the basic method of turning textual information into numerical data. Feature extraction is a simple way to improve natural language processing's text interpretation. Written information has to be cleaned and standardized before it can be translated into its characteristics for modeling, because machines cannot calculate it directly. Because they are easily processed by computers, the words are represented mathematically. Countvectorizer, Word Embeddings, and TF-IDF Vectorizer are just a few of the feature extraction techniques available. Procedures (F.) Since this study makes use of the Python programming language, one of its libraries, the NLTK (Natural Language Toolkit), has the ability to serve a variety of purposes. The library uses the most famous and influential classifier, the Naïve Bayes Classifier, among its many classification methods. Using the Bayes theorem as its foundation, the Naive Bayes method applies supervised learning to the task of solving classification problems. Text classification using a big training set is its primary use. Because it is a probabilistic classifier, it uses the probability of an object's presence to generate predictions. G. Architecture of the System Fig. 2 shows the system architecture for collecting and acquiring data on suicide attempts; this data is then split into a training dataset and a validation dataset, which are given simultaneously for the purpose of training a classification model. After training the dataset, the trained model may be used to analyze, extract, and classify different tweets. An example would be feeding the trained model a subset of the available data. We can test the model to check whether it gets the intended conclusion since the result of the input data posted on Twitter is known. What is intended here is an evaluation of the trained model's performance. The result would be postings that may be considered suicidal, or it wouldn't.

IMPLEMENTATION

There are now more online social networking venues than ever before, where everyone may post their views—good or bad. Posts that encourage hate speech self-harm, depression, or suicide thoughts demonstrate how open a person is to expressing their views online, even when positive ideas may have a smaller audience. Twitter, a social networking service that allows users to quickly express their thoughts on a range of subjects, is a good example of this. Connectivity to various regions of the globe shows that Twitter's communication reach is worldwide.

This research primarily aims to examine and categorize online comments or tweets as suicidal or not, as well as the propagation of such negative views. Section A. Python The programming language known as Python is interpreted, high-level, object-oriented, and dynamically semantic. Rapid application building and use as a glue language or scripting language to bring existing components together are made possible by its high-level integrated data structures, type dynamics, and dynamic binding. Python's short syntax prioritizes readability and usability, which lowers the cost of programming maintenance. Program modularity and code reuse are fostered by using Python's modules and packages. For each popular operating system, you may download the Python interpreter and its extensive standard library in either source or binary format, and they're both free to use. Many programmers find themselves enamored with Python due to its enhanced productivity. Because there is no compilation process step, the edit-test-debug cycle is

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

very fast. Because segmentation failures in Python programs are almost never caused by bugs or incorrect input, they are quite straightforward to identify. Instead, the language interpreter will issue an error message whenever an error is found. A stack trace is produced by the interpreter if the code fails to catch the error. You may create breakpoints, test arbitrary expressions, analyze both global and local parts, navigate line-by-line code, and more using a source-level debugger. One use of introspection is the debugger, which was built in Python. Because of the brief edit-test-debug cycle, however, adding a few print instructions to the code base is the fastest method to debug a program. Python is often used for a multitude of purposes because to its versatility. One common use is in the creation of online applications using frameworks such as Turbo gears, Django, and Zope. Manage systems with brief scripts • Using GUI toolkits to develop desktop applications, such as Tkinter or wxPython, or, more recently, Windows Forms and Iron Python • Creating Windows apps using Py2exe to generate independent programs and the Pywin32 extension for completed Windows integration. • Use of the R packages Scipy and Matplotlib in academic research. Python will be written for this project using a text editor. Coding with Python is best accomplished using an IDE such as Thonny, Pycharm, Netbeans, or Eclipse. These integrated development environments are great for managing a massive number of Python files.

B. Notebook for Jupiter the Jupiter Notebook is a free and open-source web software that allows users to create and share documents containing text, pictures, equations, and code that is updated in real-time. Some examples include data transformation and purification, data visualization, machine learning, computational statistics, and other related areas. Jupiter notebook is a useful tool for developing and enhancing Python applications used for data analysis. Instead of building and rewriting a whole program, you may generate smaller pieces of code and run them individually. Then, you may go back to the same window, make the change, and then run the application again if you want to make a modification. The foundation of Jupiter Notebook is I Python, an interactive terminal window execution environment for Python that follows the Read-Eval-Print-Loop (REPL) paradigm. The calculation-doing IPython Kernel is in constant communication with the user-facing Jupiter Notebook. Jupiter Notebook may now communicate in more than one language. With the help of Jupiter Notebooks, IPython can save things like markdown comments, original source code, and results. 1) Sign up for a Jupiter platform: Launch Jupiter Notebook by opening your terminal and navigating to the directory where you want to save it. After you execute the command Jupiter notebook, the application will start a local server at localhost: 8888 (or another port you choose). If it doesn't, go to the URL it gives you. If it does, the Jupiter Notebook user interface should pop up in your browser. Notebooks are assigned a unique token since the application uses pre-built Docker containers to provide them separate routes. To exit the terminal and turn off the server and kernel, press control-C twice. 2) Jupiter Interface: You can now see all of the files in the directory you're in the Jupiter Notebook interface. As a visual hint, each Jupiter Notebook has a representation next to its name. In the list of files, locate the Jupiter Notebook that you want to open in the current directory. Then, click on it to open it. Select Notebook - Python 2 from the new menu to begin creating a new notebook. If you would like to utilize another Jupiter Notebook on your computer, you may access them by clicking the Upload button. If a notebook is currently in use, its icon will be green; if it is not, its icon will be gray. To get a rundown of all running notebooks, go to the Running tab. thirdly, inside the notepad: The presence of a cell is the primary identifier of a new Jupiter notebook. Cells are the areas where the code is written. Additionally, it is used as a foundation for notebooks.

Fig. 3. Various cells displayed in the jupyter notebook when opened

As shown in Fig. 3, when the project is launched in the appropriate way, Jupiter notebook displays many cells. To execute the code, select the cell by clicking on it and then press SHIFT+ENTER or the play button in the toolbar up above. Under the Cell menu, you'll also discover a number of choices for running cells, such as running all cells at once or just one cell. Part VI: Findings and Prompts When you choose the Restart and Run all option in kernel mode in the Jupiter Notebook, all of the cells in the notebook begin running simultaneously. Finding out whether a tweet is suicidal is the main goal of the research. Figure 4 displays the result of executing the code in the notebook in a certain sequence to get the required output from several cells containing the source code.

Fig. 4. Result of the cell as a suicidal tweet or not

CONCLUSION

This study introduced a Machine Learning technique to identify tweets that may be suicidal or not. We used Natural Language Processing (NLP) and other ML frameworks to make predictions about the results. But these methods can only identify suicide-related tweets that have already been posted, so they can't predict when suicidal thoughts may emerge. Ultimately, this research aims to enable a notification feature so that it can collaborate with government agencies dealing with suicide prevention and to expand this feature to other social media platforms outside of Twitter. It will also attempt to predict mild symptoms of suicidal thoughts, such as anxiety, depression, fear, low self-esteem, etc.

REFERENCES

[1]. Arwa Alshamsi, Reem Bayari, Said Salloum, ”Sentiment Analysis in English Texts”, 2021

[2]. G. Gautam, D. Yadav, “Sentiment analysis of twitter data using machine learning approaches and semantic analysis,”IEEE: 437–442, 2014.

[3]. Atika Mbarek, Salma Jamoussi, Anis Charfi,and Abdelmajid Ben Hamadou,”Suicidal Profiles Detection in Twitter”,2019.

[4]. Hannah Metzler, Hubert Baginski, Thomas Niederkrotenthaler, David Garcia, ”Detecting Potentially Harmful and Protective Suicide-related Content on Twitter: A Machine Learning Approach”

[5]. Md Taufiqul Haque Khan Tusar, Md. Touhidul Islam,”A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learn- ing Techniques on US Airline Twitter Data. 2021.

[6]. Feature Extraction in Natural Language Processing provided by: https://turbolab.in/feature-extraction-in-natural-language-processing- nlp/

[7]. Z.J. and G. Xiaolini, “Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis”,2017 https://www.semanticscholar.org/paper/A-ComparativeStudy-of-Sentiment-Analysis-Using-NLP-TusarIslam/393aea1268bdecd80f6346a3f362946546de9b0f?sort=releva ncepdf=true

[8]. Public Health Agency of Canada. (2019) Suicide in Canada: Key statistics. Retrieved from: https://www.canada.ca/en/public-health/services/publications/healthy-living/suicide-canada-key-statistics- infographic.html.

[9]. Lee, S. Y. Kwon, Y. Twitter as a place where people meet to make suicide pacts. Public Health 159, 21–26 (2018).

[10]. O’Dea, B., Larsen, M. E., Batterham, P. J., Calear, A. L. Christensen, H. A linguistic analysis of Suicide-related twitter posts. Crisis 38, 319–329 (2017).

[11]. Van Orden, K. A. et al. The interpersonal theory of suicide. Psychol. Rev. 117, 575–600 (2010).

[12]. Isometsa, E. T. et al. The last appointment before suicide: is suicide intent communicated? Am. J. Psychiatry 152, 919–922 (1995).

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

[13]. Spates, K., Ye, X. Johnson, A. "I just might kill myself": Suicide expressions on twitter. Death Stud. 44, 189–194 (2020).

[14]. Marchant, A. et al. A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: the good, the bad and the unknown. PLoS ONE 12, e0181722 (2017).