**Article**

ANALYSING SUICIDE IDEATION VIA TWEETS USING NATURAL LANGUAGE PRCESSING (NLP)

**Authors :- Abhishek Singh\*, Aditya Pratap Singh\*, Avanish Singh\*, Manas Gupta\***

**\* B.Tech IInd-year (CSE- Artificial Intelligence and Machine Learning) G.L.A University, Mathura, Uttar Pradesh**

**Abstract :** Suicide ideation expressed on social media has a linguistic influence. Many at-risk people use social forum platforms to talk about their issues or find out about similar tasks. The main goal of our research is to present ongoing work on automatic suicidal post recognition. Using deep learning and machine learning-based classification methods applied to Twitter social media, we address the early detection of suicidal ideation. We use natural language processing to determine whether or not the person intends to commit suicide for this purpose. In addition, our findings show that deep learning architectures can be used to create an effective model for assessing suicide risk in a variety of text classification tasks.

**Keywords:** suicide ideation; early suicide detection; deep earning; machine learning; twitter social media;



1. **INTRODUCTION**

*Suicide has long been one of the leading causes of death around the world. Suicide took the lives of 828,000 people worldwide in 2015, up from 712,000 in 1990, according to Wikipedia. Suicide is also the tenth leading cause of death in the world. There's also mounting evidence that the Internet and social media may have an effect on suicidal actions. In recent years, social media has evolved into a strong "window" into the mental health and well-being of its users, who are mostly young people. It allows anonymous participation in various cyber communities in order to provide a forum for public debate on socially stigmatized topics.*

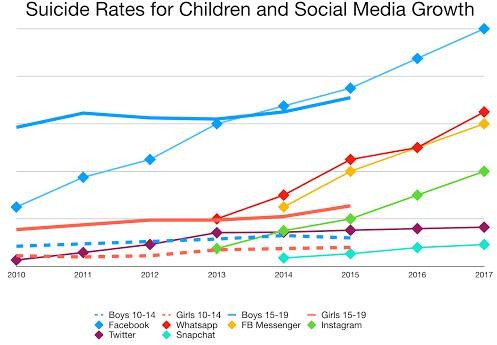
*As a result, any written suicidal sign should be regarded as concerning, and an individual should be questioned about the nature of personal thoughts. According to some sources, social media text such as blog posts, forum comments, tweets, and other online notes are typically documented throughout the present and well preserved. It can reduce and misleading text interpretations provided by a retrospective analysis when compared to an offline text.*

*In computational linguistics, social media, with its mental health-related forums, has emerged as a new research field. It offers a valuable research forum for the advancement of new technical methods and changes that can lead to new ways of detecting suicide and reducing suicide risk. It may be useful as a point of interference. Deep learning techniques have also made significant progress in the fields of computer vision and pattern recognition, in addition to conventional text classification approaches. Traditional machine learning approaches rely heavily on time-consuming and sometimes incomplete handcrafted features, but neural networks based on dense vector representations can outperform them on a variety of NLP tasks. Word embedding and deep neural networks are outperforming more conventional machine learning systems in suicide risk assessments, demonstrating their growing performance. The most important*

*The aim of our research is to share our understanding of suicide ideation in tweets through data analysis and the use of an efficient NLP algorithm.*

# Some Related Work

*In recent years, a large number of studies have been conducted to demonstrate the impact of social media on suicidal ideation. Choudhury devised a mathematical method based on a score matching model to derive some distinct markers for detecting the transition from mental health discourse to suicidal ideation. According to the writers, three psychological phases will accompany this transition: thought, ambivalence, and decision-making.*



*Anxiety, hopelessness, and depression are all part of the first step. The second stage is associated with a loss of self-esteem and social stability. Aggression and a suicide commitment plan accompany the third step.*

*Suicidal behaviors are more effectively recognized as regular language variations of social media text are identified. It is often aided by the application of various machine learning methods to various NLP techniques. Desmet used binary Support Vector Machine (SVM) classifiers to create a suicide note analysis tool for detecting suicide ideation. A psychological lexicon based on a Chinese sentiment dictionary was developed by Huang (Hownet). He used the SVM technique to find a classification for a real-time suicidal ideation warning device on Chinese social media platform Weibo. Braithwaite showed that machine learning algorithms are effective at distinguishing between individuals who are and are not suicidal. Sueki conducted research into the suicidal intent of Japanese Twitter users in their twenties, concluding that language framing is critical for recognising suicidal markers inside the text For example, the word "want to suicide" is more often associated with a lifelong suicidal intent than the expression "want to die."*

*CNN neural networks with convolutional, nonlinear, and pooling layers have been successfully applied to a wide variety of NLP tasks in the recent past, and have proved to outperform conventional NLP approaches. There have been numerous attempts to establish a hybrid paradigm for coherent combinations of CNNs and RNNs to apply the merits of both. In this paper, we suggest a paradigm for recognising suicidal ideation in social media that is built on a Natural Language Processing (NLP) model.*

# *Dataset*

*To detect suicidal ideation, we use a dataset of Tweets and comments in which users can express themselves through text posts, links, or voting mechanism posts to train our classification models. They interact with one another by comment threads that are attached to each message. The dataset used in our project is a list of suicide-indicative and non-suicidal posts that can be found on kaggle.com. We have made some improvements to the original dataset, such as adding labels and annotations ourselves, to improve the model's accuracy.*

*Our dataset is made up of 1590 separate tweets from a vast number of people who are committed to helping people who are at risk. Non-suicidal tweets are taken from subjects related to families and friends.*



One question that both of us have is, "What does a suicidal post look like?" What is the *format of a non-suicidal post? So, as you can see, figs. 2 and 3 depict Suicidal and Non-Suicidal messages, respectively.*



Fig 2 Example of Suicidal post Fig. 3 Example of Non Suicidal post

*And in fig.4 we have given the illustration of dataset used to train this model :-*

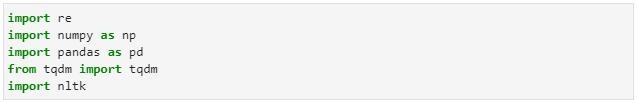


# Methodology

*The aim of this study is to improve the efficiency of language modelling and text classification in detecting suicidal ideation in tweets by using a deep learning classifier. In our experiment, we provide a technical overview of approaches that use different NLP and text to distinguish techniques.*

## Data Processing

*For the purpose of Data Processing we have to import libraries to our python program.*



And, download Package name ‘Stopwords’ from nltk library.

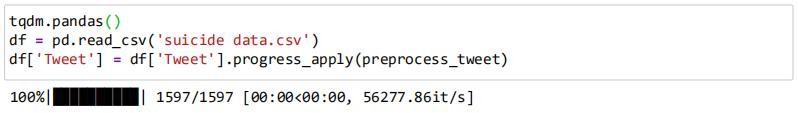


## Pre-Processing

*Pre-processing is the task of eliminating unnecessary functionality from an input text in order to enhance the consistency of a proposed system before studying word embedding. It's achieved by adding a series of filters to tweets to turn raw data into a format that learning models can comprehend. In our study, we discovered that we pre-process the dataset using the Natural language toolkit (NLTK) [12] before going on to the training level.*

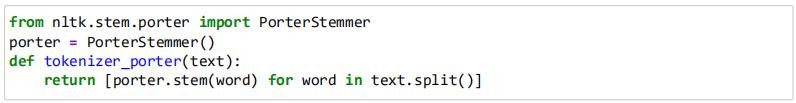
*We developed a feature that cleans text data by deleting HTML markup, retaining emoticon characters, eliminating non-word characters, and eventually converting it to lowercase. Removes all HTML markup from text files, keeps emoticon characters, removes non-word characters, and eventually converts to lowercase.*

*After defining function we have applied preprocess\_tweet function to tweets dataset to clean the data.*



## Tokenizing

*Then we have to convert the text to tokens by using the .split() method and used word stemming to convert the text to their root form.*

**

*After creating tokens we have to import the stopwords library to remove stop words in the text.*



*Testing the function on a single text.*

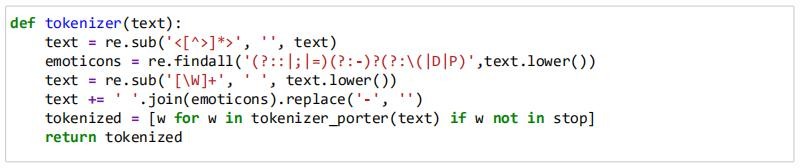


## Output



* 1. **Vectorizer**

*We chose the Hashing Vectorizer for this project because it is data-independent, which means it uses very little memory, is robust to massive datasets, and does not keep a vocabulary dictionary in memory. The Hashing Vectorizer was then given a tokenizer function.*

**

*Then, We created a Hashing Vectorizer Object.*



## Model

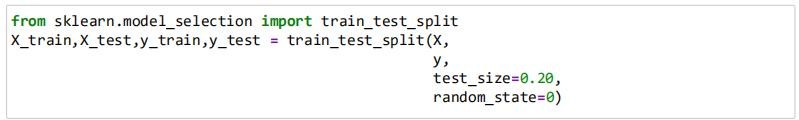
*For the model, we have used stochastic gradient descent classifier algorithm.*



## Training and Validation



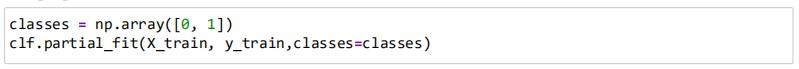
*For the model, We used 80% for training and 20% for testing*.



*Then we transformed the text data to vectors with the Hashing Vectorizer we created earlier:*



*Finally, we then fit the data to the algorithm.*

**

*And, Now we tested Accuracy of our model.*



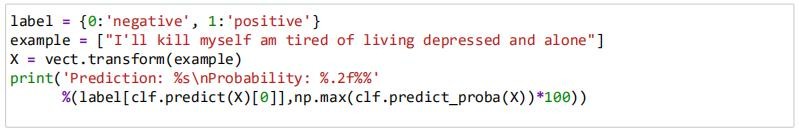
Output :-



We had an accuracy of 92% which is fair enough.

# Testing and Making predictions

We added the text “I’ll kill myself am tired of living depressed and alone” to the model.



**Output :-**



And when we used the following text “It’s such a hot day, I’d like to have ice cream and visit the park”, I got the following prediction:



The model was able to predict accurately for both cases.

## Conclusion

*Our findings suggest that applied classification methods perform reasonably well; however, the absolute importance of the metrics indicates that this is a complex challenge that requires more research. In the future, we could try to get access to a larger dataset with suicidal ideation material, as well as a new dataset with similar themes. Suicidal ideation and causes such as family climate, temperature, and so on should all be investigated. Both datasets will be collected from a variety of social media sites in order to better explain and compare our proposed hybrid model.*

**References**

1. *Articl* **Detection of Suicide Ideation in Social Media Forums Using Deep Learning (By : Michael Mesfifin Tadesse , Hongfei Lin , Bo Xu and Liang Yang )**
2. Dataset :

[https://www.kaggle.com/imeshsonu/suicideal-phrases?select=Train\_suicide1.cs](https://www.kaggle.com/imeshsonu/suicideal-phrases?select=Train_suicide1.csv) v

1. Modified dataset : [https://drive.google.com/file/d/1s1cCfmMGgMdnMEsvlHk3g6E0FbeHR8CN/v](https://drive.google.com/file/d/1s1cCfmMGgMdnMEsvlHk3g6E0FbeHR8CN/view?usp=sharing) iew?usp=sharing