iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IX **Month of publication:** September 2020

DOI: https://doi.org/10.22214/ijraset.2020.31392

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

Sentiment Analysis to Detect Mental Depression Based on Twitter Data

Vishal Tyagi[1], Shraddha Singh[2]
[1, 2]*B.tech, Student, Computer Science, Raj Kumar Goel Institute of Technology, Uttar Pradesh, India*

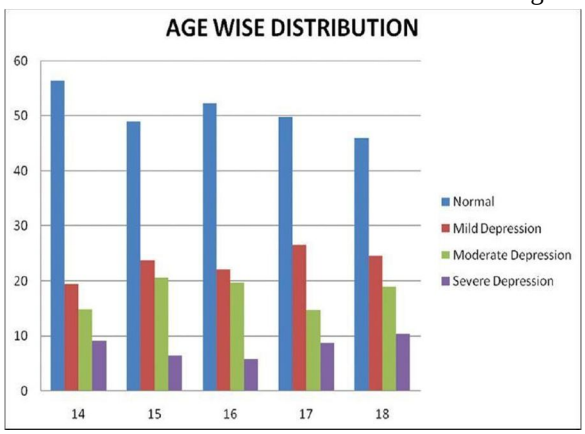
Abstract: ***The objective of this briefing is to present an overview about the use of sentiment analysis to detect mental depression. Mental illness is among the most prevalent yet overlooked issues. According to the World Health Organization, around 20% of the child and adolescent population and 23% of the total human population have one or more mental illnesses, making neuropsychiatric disorders the major cause of disability all over the world [1]. Among all, Depression is among the most prevalent mental health disorders. More than 300 million individuals or 4.4% of the world's population is estimated to suffer from depression [2]. However, its prevalence varies by the WHO region and gender, from a minimum of 2.6% among the Western Pacific males to a maximum of 5.9% among the African females.***
"Machine learning is a core, transformative way by which we're rethinking everything we're doing." - Google CEO Sundar Pichai

I. BACKGROUND

Social media is frequently used by youth to share their health and mental issues. Therefore, social media has become a major online resource to study the language used to express issues such as depression and self- harm which can help to identify individuals at risk of harm. Furthermore, depression and suicide are generally closely related especially that depression is the most common symptoms associated with self-harm acts such as suicide. In this paper, the application of sentiment analysis to depression detection is presented and discussed. Moreover, a preliminary design of an integrated system for depression detection that includes sentiment analysis techniques is proposed.

II. PROBLEM DESCRIPTION

Social Media platforms are becoming an integral part of people's life. The applications such as Facebook, Twitter, Instagram and alike not only host the written and multimedia contents but also offer their users to express their feelings, emotions and sentiments about a topic, subject or an issue online. On one hand, this is great for users of social networking site to openly and freely contribute and respond to any topic online; on the other hand, it creates opportunities for people working in the health sector to get insight of what might be happening at the mental state of someone who reacted to a topic in a specific manner. In order to provide such insight, machine learning techniques could potentially offer some unique features that can assist in examining the unique patterns hidden in online communication and process them to reveal the mental state (such as 'happiness', 'sadness', 'anger', 'anxiety', depression) among social networks' users. Informed societies are new beneficiaries of big data.



What starts as social online discourse could hold prime value for identifying pertinent risks and opportunities. Researchers are using information garnered from social media to provide insights to explain human behaviour.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

III. NATURAL LANGUAGE PROCESSING

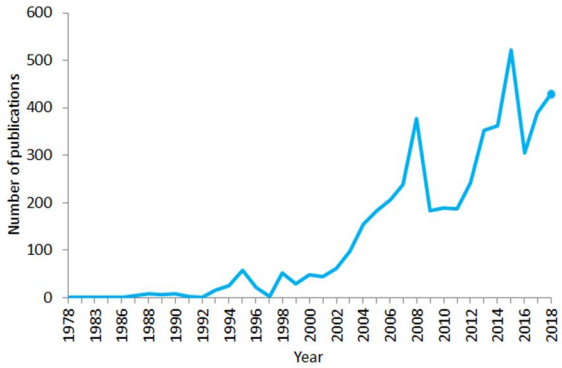
Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.

A. Use Cases of NLP

In simple terms, NLP represents the automatic handling of natural human language like speech or text, and although the concept itself is fascinating, the real value behind this technology comes from the use cases.NLP can help you with lots of tasks and the fields of application just seem to increase on a daily basis. Let's mention some examples:

1) NLP enables the recognition and prediction of diseases based on electronic health records and a patient's own speech. This capability is being explored in health conditions that go from cardiovascular diseases to depression and even schizophrenia. For example, Amazon Comprehend Medical is a service that uses NLP to extract these conditions, medications and treatment outcomes from patient notes, clinical trial reports and other electronic health records.

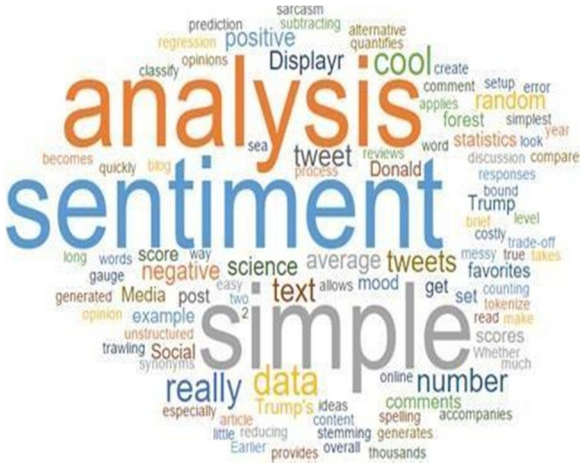
2) Organizations can determine what customers are saying about a service or product by identifying and extracting information in sources like social media. This sentiment analysis can provide a lot of information about customers' choices and their decision drivers.



Above are the number of publications containing the sentence "natural language processing" in PubMed in the period 1978–2018. As of 2018, PubMed comprised more than 29 million citations for biomedical literature.

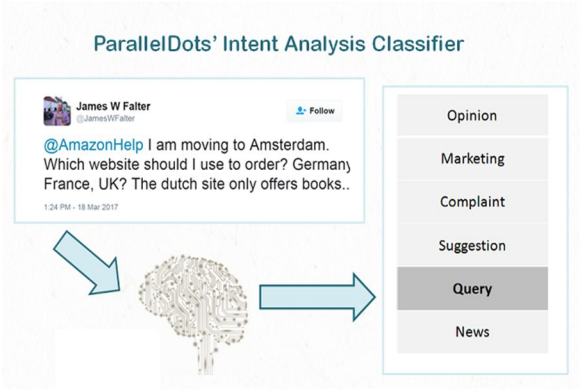
IV. WHAT IS SENTIMENT ANALYSIS?

Sentiment analysis is the interpretation and classification of emotions within voice and text data using text analysis techniques, allowing businesses to identify customer sentiment toward products, brands or services in online conversations and feedback. The task of sentiment analysis typically involves taking a piece of text, whether it's a sentence, a comment or an entire document and returning a "score" that measures how positive or negative the text is.



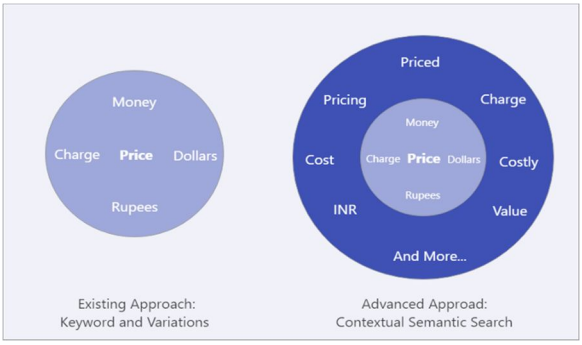
A. *Intent Analysis*

Intent analysis steps up the game by analyzing the user's intention behind a message and identifying whether it relates an opinion, news, marketing, complaint, suggestion, appreciation or query.



B. *Contextual Semantic Search(CSS)*

The way CSS works is that it takes thousands of messages and a concept (like **Price**) as input and filters all the messages that closely match with the given concept. The graphic shown below demonstrates how CSS represents a major improvement over existing methods used by the industry.



A conventional approach for filtering all Price related messages is to do a keyword search on Price and other closely related words like (pricing, charge, $, paid). This method however is not very effective as it is almost impossible to think of all the relevant keywords and their variants that represent a particular concept. CSS on the other hand just takes the name of the concept (Price) as input and filters all the contextually similar even where the obvious variants of the concept keyword are not mentioned.

C. *Sentiment Library*

Much in the way your brain remembers the descriptive words you encounter over your lifetime and their relative "sentiment weight", a basic sentiment analysis system draws on a sentiment library to understand the sentiment-bearing phrases it encounters.

Sentiment libraries are very large collections of adjectives (good, wonderful, awful, horrible) and phrases (good game, wonderful story, awful performance, horrible show) that have been hand-scored by human coders. This manual sentiment scoring is a tricky process, because everyone involved needs to reach some agreement on how strong or weak each score should be relative to the other scores. If one person gives "bad" a sentiment score of -0.5, but another person gives "awful" the same score, your sentiment analysis system will conclude that both words are equally negative.

What's more, a multilingual sentiment analysis engine must maintain unique libraries for each language it supports. And each of these libraries must be maintained constantly: scores tweaked, new phrases added, irrelevant phrases removed.

D. *Rules-based Sentiment Analysis Systems*

Once the sentiment libraries are prepared, software engineers write a series of guidelines ("rules") to help the computer evaluate the sentiment expressed towards a particular entity (noun or pronoun) based on its nearness to known positive and negative words (adjectives and adverbs).

E. Drawbacks of rules-based sentiment analysis

The simplicity of rules-based sentiment analysis makes it a good option for basic document-level sentiment scoring of predictable text documents, such as limited-scope survey responses. However, a purely rules-based sentiment analysis system has many drawbacks that negate most of these advantages. A rules-based system must contain a rule for every word combination in its sentiment library. Creating and maintaining these rules requires tedious manual labor. And in the end, strict rules can't hope to keep up with the evolution of natural human language. Instant messaging has butchered the traditional rules of grammar, and no ruleset can account for every abbreviation, acronym, double-meaning and misspelling that may appear in any given text document.

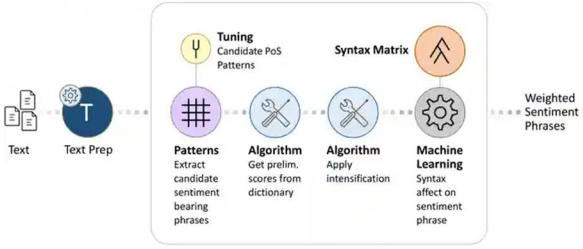
In addition, a rules-based system that fails to consider negators and intensifiers is inherently naïve, as we've seen. Out of context, a document-level sentiment score can lead you to draw false conclusions. Lastly, a purely rules-based sentiment analysis system is very delicate. When something new pops up in a text document that the rules don't account for, the system can't assign a score. In some cases, the entire program will break down and require an engineer to painstakingly find and fix the problem with a new rule.

In the end, anyone who requires nuanced analytics, or who can't deal with ruleset maintenance, should look for a tool that also leverages machine learning.

F. Hybrid Sentiment Analysis System

Hybrid sentiment analysis systems combine machine learning with traditional rules to make up for the deficiencies of each approach. Rules-based sentiment analysis, for example, can be an effective way to build a foundation for PoS tagging and sentiment analysis. But as we've seen, these rulesets quickly grow to become unmanageable. This is where machine learning can step in to shoulder the load of complex natural language processing tasks, such as understanding double-meanings.

Most hybrid sentiment analysis systems combine machine learning with software rules across the entire text analytics function stack, from low-level tokenization and syntax analysis all the way up to the highest-levels of sentiment analysis.



V. METHODOLOGY

The dataset employed is a subset of Sentiment140 dataset and used Google Colaboratory as the platform for the purpose of coding. Our methodology involves use of Naïve Bayes as a classifier for the tweets.

VI. MODEL SELECTION

The most exciting phase in building any machine learning model is selection of algorithms. We can use more than one kind of data mining techniques to large datasets. But, at high level all those different algorithms can be classified in two groups: supervised learning and unsupervised learning. In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B(Benign). So, Classification algorithm of supervised learning is applied to it. We have Naïve Bayes as the algorithm for our analysis. Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. It is called 'Naive' because of the naive assumption that the X's are independent of each other. Regardless of its name, it's a powerful formula.

VII. DATA PREPROCESSING AND EXPLORATORY DATA ANALYSIS

Data preprocessing is an important part in any data analysis task, because, during the data gathering process, there are many factors that may lead to irrelevant and out-of-range data, missing values etc. Analyzing data which has not been carefully screened for such problems can produce misleading results. Thus, the quality of data and its representation is first and foremost before running any analysis.

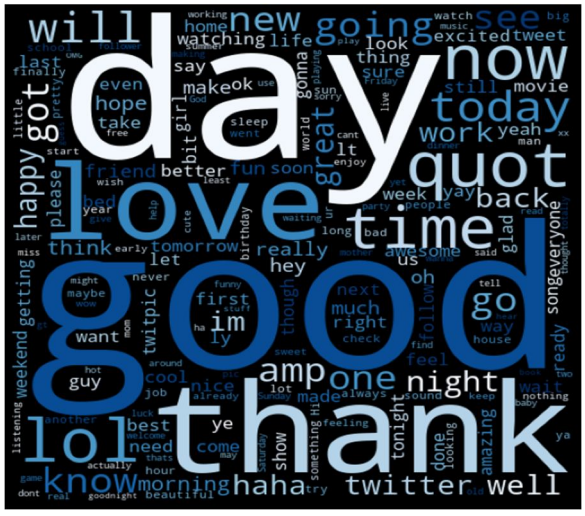
If, in data, there is too much irrelevant information or noisy and unreliable data present, then the knowledge discovery and model creation will be more difficult. So, to solve this problem, data should be prepared and filtered. And this process takes a considerable amount of processing time.

Data preprocessing includes binarization, categorization, standardization, cleaning, feature extraction etc. In our work, we used only categorization technique, because the dataset used is already good for classification problems.

Once you have got yourself a nice, cleaned dataset, the next step is Exploratory Data Analysis (EDA). EDA is the process of figuring out what the data can tell us and we use EDA to find patterns,relationships or anomalies to inform our subsequent analysis. While there are an overwhelming number of methods to use in EDA, in case of NLP or Natural Language Processing, the most significant is the wordcloud.

A wordcloud containing most frequently used words in depressive tweets are visualized.

The words that caught attention in these tweets are anxiety, depression, help, treatment, symptom, suffer, hard, issue, bipolar, suffering, sleep, sad, suicide, struggle, depressed, disorder, severe, stress, feeling, crying, cannabis, war, drug, death, emotional, food, illness, hate, need, problem, eating.All these words in tweets are characteristics of depression.



The most common keywords used in positive tweets are: Day happy, friend, morning, glad, better, start, birthday, good, love, excited, hope, great, best awesome, beautiful, cute, enjoy, friday, yay, pretty, playing, haha, cool, nice, home, funny, amazing, welcome.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

VIII. EVALUATION OF MODEL

A. *Precision*

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Precision is defined as the number of true positives divided by the sum of true positives and false positives.

B. *Recall*

The precise definition of recall is the number of true positives divided by the sum of true positives and false negatives.

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

C. *F1 Score*

The F-score, also called the F1 score or F-Measure, is a measure of a test's accuracy. The F-score is defined as the weighted harmonic mean of the test's precision and recall. This score is calculated according to:

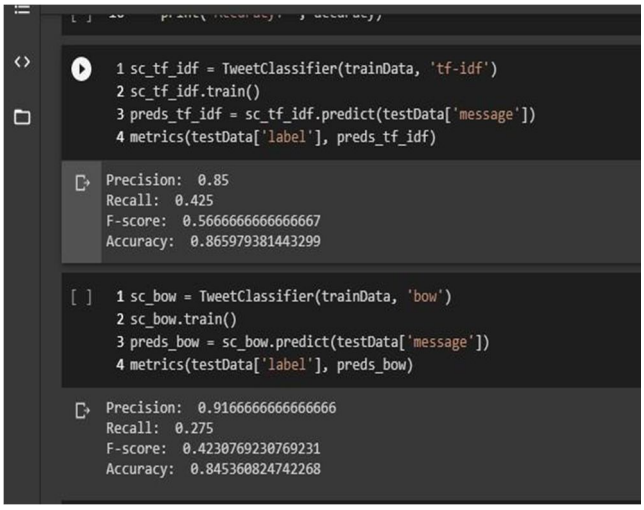
$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precisior}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precisior} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

D. *Accuracy*

Classification accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples. It works well only if there are an equal number of samples belonging to each class.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

In the proposed analysis, following results are obtained for above evaluation criteria:



IX. CONCLUSION

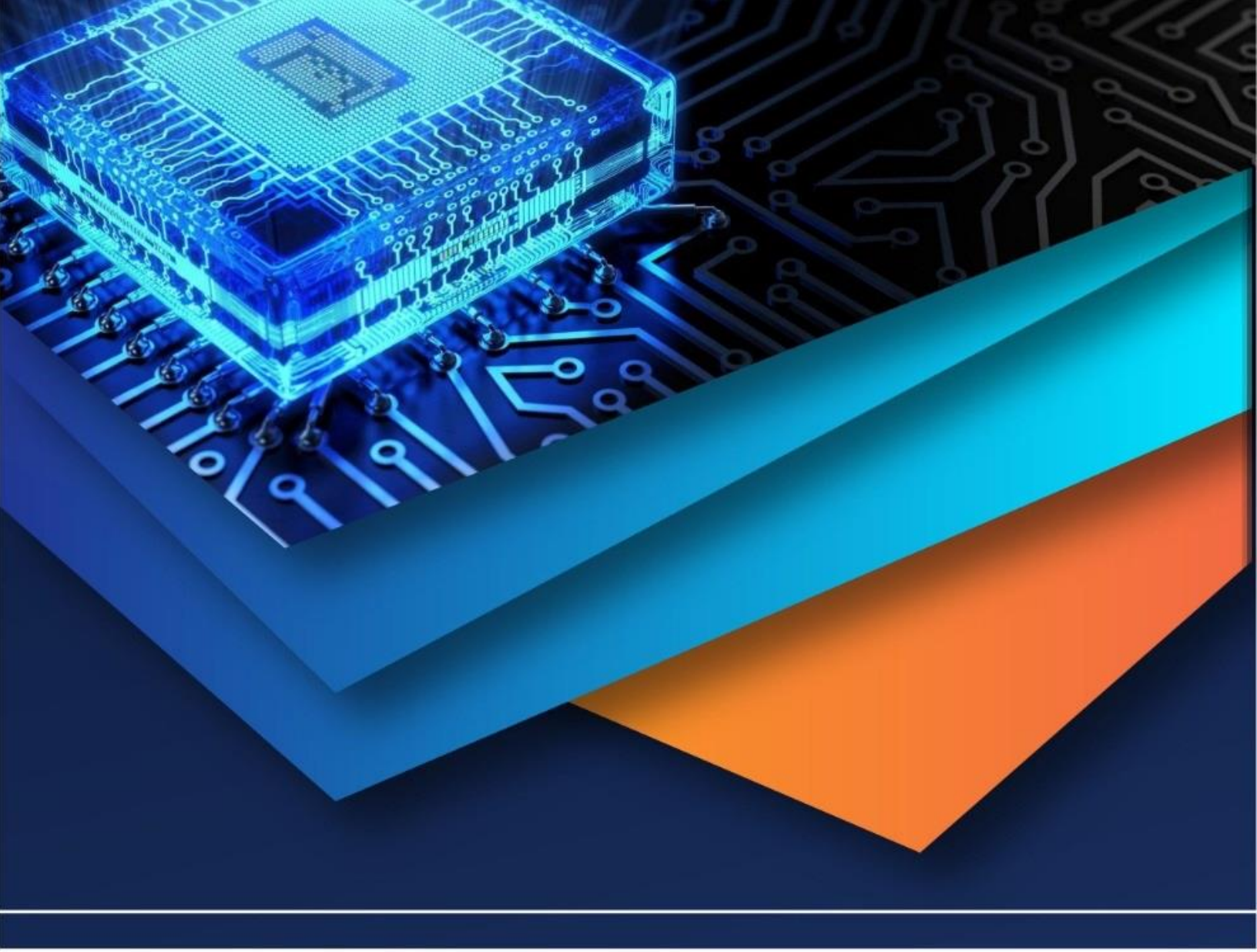
The above briefing proposes a system to detect depression from one of the most used social media platforms i.e., twitter. The people who are too scared to be talking with a therapist generally prefer putting their feelings up on social media boards.The model filters out possible slang nad abrupt keywords out of the dataset, making it useful for further analysis.Models proposed in the briefing shows promising results, making it a viable alternative for a therapist or psychiatrist to pre-assess their patients, giving them better judgement of their present condition. This work ensures that anyone suffering from depression can ensure their privacy and at the same time get the best care possible.

X. FUTURE WORKS

This is only a tiny step towards improving the lives of those suffering from depression. There are some scope for this project firstly this can be developed in multiple languages. Secondly this can be implemented using a number of additional social media platforms. In addition to this, this technology can be used to treat multiple mental illnesses such as PTSD, multiple personality disorder. Finally, some more improvement can be added to the user interface. This will provide the physicians with a better way to diagnose and cure patients with depression.

REFERENCES

[1] Dictionary.com, “Depression” (2016).
[2] Myvocabulary.com, “Depression Vocabulary Word List” (2016). https://myvocabulary.com/word-list/depressioon vocabulary/
[3] WHO, “Mental health : suicide data” (2016) http://www.who.int/mental_health/ prevention/suicide/suicideprevent/en/
[4] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, New York, NY, USA, pp. 3187–3196. ACM (2015)Google Scholar
[5] O’Dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., Christensen, H.: Detecting suicidality on twitter. Internet Interventions 2(2), 183–188 (2015) CrossRefGoogle Scholar



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY