iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.51233

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: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Cyber Bullying Detection on Social Media Network

Chindhuja P[1], Darshini K[2], Haritha M[3], Kowsalya J[4]
Department of Computer Science and Engineering, School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India.

Abstract: ***Cyberbullying has grown more prevalent on social networking sites. Cyberbullying has resulted in a significant increase in mental health issues, particularly among the younger population. Self-esteem and mental health difficulties will affect a whole generation of young adults unless action is made to combat cyberbullying. Many classical machine learning methods have already been used for the automated identification of cyberbullying on social media. Since the advent of social media platforms about 20 years ago, there haven't been many effective ways to stop social bullying, and it has recently grown to be one of the most concerning problems. In this project, we create an AI-based system for detecting cyberbullying and investigate ways to identify fraudulent profiles and abusive speech on social media while separating them from ordinary vulgarity. Applying supervised classification techniques to a manually annotated open-source dataset, we seek to create lexical baselines for this job.***
Keywords: Cyberbullying, Social media, Natural Language Processing (NLP), Swift algorithm.

I. INTRODUCTION

Social media has become an integral element of human existence. Because of the increased usage of social media, cyberbullying has grown more prevalent on social networking sites. The use of the web and social media typically leads to the sharing, receiving, and posting of unpleasant, abusive, misleading, or malicious content about another person, which is considered as cyberbullying. It has led to decreased self-esteem and increased suicide thoughts. The era of social networking began with the creation of the internet. No one could have predicted that the internet will one day house a plethora of great services such as social networking. Online apps and social networking websites have now become an inseparable part of one's life. Many people of all ages spend hours each day on such websites. Despite the fact that social media allows individuals to connect emotionally, it also brings with it significant risks such as cyber-attacks and cyberbullying. While social media networks provide excellent communication platforms, they also make young people more vulnerable to hazardous circumstances online. Because of the large number of active users on social media networks, cyberbullying is a worldwide issue.

For instance, with the proliferation of social media sites (e.g., Facebook, Twitter) being available online, 'hate groups' have become a viable tactic to bullying. However, social media sites have started to take measures against the emergence of hate organizations. For instance, when forming a facebook group, a warning near the bottom of the page claims, "Note: groups which abuse a specific individual or body of citizens (e.g., racist, sexist, and perhaps other hate groups) will not be permitted."

The research reveals that cyberbullying on social media is becoming more prevalent by the day. According to recent studies, cyberbullying is becoming more prevalent among children. Successful prevention is dependent on accurate identification of potentially hazardous messages, and the information overload on the Internet necessitates the use of intelligent technologies to detect possible threats automatically.

II. LITERATURE SURVEY

In their study " Cyberbullying Detection using Pre-Trained BERT Model ," D. Chauhan et al., introduces a new and novel discipline that uses the BERT model with a single linear neural network layer on top as a classifier to provide a novel method for detecting cyberbullying on social media sites. The Form Spring forum and Wikipedia datasets are used to train and assess the model. In comparison to the previously used models, the proposed model's performance accuracy for the Form spring dataset was 98% and for the Wikipedia dataset it was 96%. Due to the Wikipedia dataset's vast size, the suggested approach produced better results without the requirement for oversampling than it did for the Form spring dataset. [3]

The purpose of Rachel E. Trana et al., paper “Fighting Cyberbullying: An Analysis of Algorithms Used to Detect Harassing Text Found on YouTube” is to study and explain how to reduce exceptional occurrences involving text retrieved from picture memes using a machine learning model. Around 19,000 YouTube text views may be found in the author's database. This study evaluates the performance of the three machine learning methods used on the YouTube database: Uninformed Bayes, Support Vector Machine, and convolutional neural network. The results are compared with the pre-existing Form datasets. They searched the YouTube database's subcategories for algorithms that detect online bullying. [4]

"Towards the design of a platform for abuse detection in OSNs using multimedia data analysis," by P. Leroux et al. considers the intention to reduce exceptional occurrences involving text, retrieved from picture memes using a machine learning model. A database that the author has created contains around 19,000 text views from YouTube. This study compares the outcomes with the pre-existing Form datasets and examines the performance of the three machine learning algorithms utilized on the YouTube database: Uninformed Bayes, Support Vector Machine, and convolutional neural network. They looked at sub-categories of the YouTube database for algorithms for cyberbullying on the internet. [5]

"Automatic detection of cyberbullying on social networks based on bullying features," by Rui Zhao et al. discusses a representation learning system designed specifically for detecting cyberbullying.

They develop a list of pre-defined insulting words using word embeddings, give each word a different composition, and procure bullying attributes.

Combining the bullying traits with lexicons and semantic features they are feed to the linear SVM classifier. An experimental investigation using a twitter dataset evaluates their strategy to a number of standard text representation learning methods and cyberbullying detection approach. This is when it has been demonstrated that this strategy produces higher results. [6]

III. METHODOLOGY

Due to its prevalence and rapid growth made possible by information technologies, cyberbullying is a significant issue for today's society and has a major adverse effect on the victims.

Therefore, to minimize the impact on the victims, early detection of cyberbullying via social media is essential.

Text, user demographics, and social network traits are three types of information that are frequently employed in cyberbullying detection.

Here on the proposed approach, we seek to investigate several methodologies that consider both the time needed for detection as well as the appropriate detection of cyberbullying in social networks. The first procedure includes a trained model file to identify derogatory terms.

The next method is to identify fraudulent profile usage, by comparing the primary user image with other accounts using the Swift algorithm. Once the fake usage is identified, the system bans that specific website.

The machine learning model has five phases :

1) Defining Architecture
2) Compiling the Model
3) Fitting the Model
4) Evaluating and Making Predictions
5) Deploying The Model

This model enables the system to achieve higher accuracy levels.

A. Algorithm

a) Step 1: Preprocessing of the data
b) Step 2: Applying NLP on the training set
c) Step 3: Detecting the fake profile using Swift algorithm
d) Step 4: Predicting the outcome of the test
e) Step 5: Verify the result's accuracy
f) Step 6: Visualizing the results of the test set

IV. SYSTEM MODEL

The system uses a dataset of approximately 20001 that is preprocessed, later to which Natural Language Processing algorithm and OpenCV, to the abuse dataset that is categorized and the input image is contrasted with the trained model file.

Finally, the profile photo is analyzed with the database using the Swift algorithm to distinguish between false profile users and the genuine users and the model file is created and the classification result is presented.

The proposed system has the following modules:

1) *Module 1:* Import the dataset.
2) *Module 2:* Perform dataset pre-processing.
3) *Module 3:* Train the dataset and create the model file.
4) *Module 4:* To classify the category using Natural Language Processing-NLP machine learning technique and to detect the fake profile using Swift algorithm.

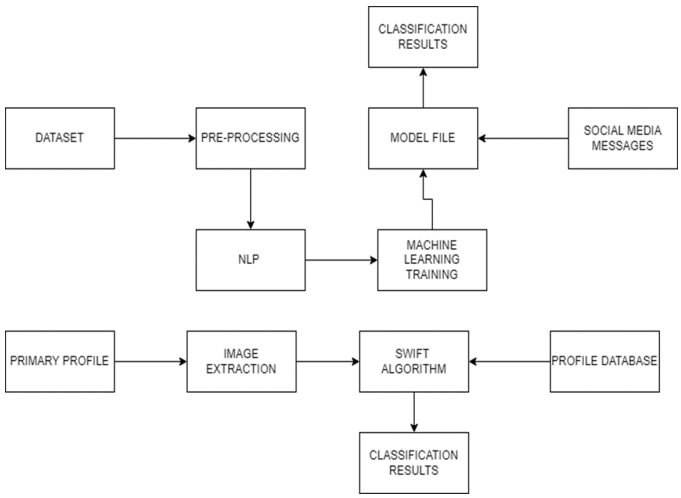


Fig 1: Workflow

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
import pickle
data = pd.read_csv('/content/dataset.csv')
data.head()
data.shape
X = data.iloc[:,:-1]
X.head()
y = data.iloc[:,-1]
y.head()
data['target'].value_counts()
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=1)
sns.countplot(x='target',data=data)
plt.show()
X_train.shape
X_train.head()
y_test.shape
y_test.head()


from sklearn.neighbors import KNeighborsClassifier

model = KNeighborsClassifier(n_neighbors=3)
model.fit(X_train,y_train)
filename = 'model.sav'
pickle.dump(model, open(filename, 'wb'))
y_pred = model.predict(X_test)
```

Fig 2: Working Source code

A. Dataset cleaning and preprocessing

The dataset's comprehensive description is provided below:

- It is a partially manually labeled dataset.
- 20001 total instances

The dataset has two components: tweet and label, where 0 indicates a No and 1 indicates a Yes.

1) Data Cleaning

The original set of fields in the annotation attribute were eliminated and replaced with the label values to make the following steps simpler because the dataset's fields were very straightforward to comprehend. Table 1 lists the number of occurrences for each class.

Instances	Message
Total Instances	20001
Cyber Bullying instances	7822
Non-Cyber Bullying instances	12179

Table 1 :Instances

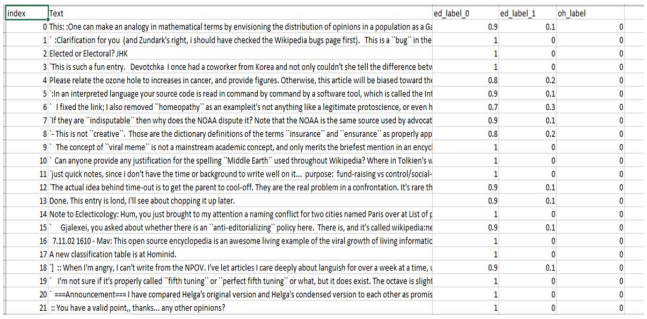
2) Preprocessing

Preprocessing data involves effectively removing undesired elements from the dataset. The dataset for this model is collected from data science website called Kaggle

Several steps are involved in developing this model to enhance the powerful data, for effective and accurate detection of cyberbullying.

The steps are:

- Acquire the dataset
- Import all the crucial libraries
- Import the dataset
- Identifying and handling the missing values
- Encoding the categorical data
- Splitting the dataset
- Feature scaling



index	Text	ed_label_0	ed_label_1	oh_label
0	This: :One can make an analogy in mathematical terms by envisioning the distribution of opinions in a population as a Ga	0.9	0.1	0
1	:Clarification for you (and Zundark's right, i should have checked the Wikipedia bugs page first). This is a ''bug'' in the	1	0	0
2	Elected or Electoral? JHK	1	0	0
3	This is such a fun entry. Devotchka I once had a coworker from Korea and not only couldn't she tell the difference betv	1	0	0
4	Please relate the ozone hole to increases in cancer, and provide figures. Otherwise, this article will be biased toward th	0.8	0.2	0
5	:In an interpreted language your source code is read in command by command by a software tool, which is called the Int	0.9	0.1	0
6	I fixed the link; I also removed ''homeopathy'' as an exampleit's not anything like a legitimate protoscience, or even h	0.7	0.3	0
7	If they are ''indisputable'' then why does the NOAA dispute it? Note that the NOAA is the same source used by advocat	0.9	0.1	0
8	- This is not ''creative''. Those are the dictionary definitions of the terms ''insurance'' and ''ensurance'' as properly app	0.8	0.2	0
9	The concept of ''viral meme'' is not a mainstream academic concept, and only merits the briefest mention in an encycl	1	0	0
10	Can anyone provide any justification for the spelling ''Middle Earth'' used throughout Wikipedia? Where in Tolkien's w	1	0	0
11	just quick notes, since i don't have the time or background to write well on it... purpose: fund-raising vs control/social-	1	0	0
12	The actual idea behind time-out is to get the parent to cool-off. They are the real problem in a confrontation. It's rare th	0.9	0.1	0
13	Done. This entry is lond, I'll see about chopping it up later.	0.9	0.1	0
14	Note to Eclecticology: Hum, you just brought to my attention a naming conflict for two cities named Paris over at List of p	1	0	0
15	Gjalexei, you asked about whether there is an ''anti-editorializing'' policy here. There is, and it's called wikipedia:ne	0.9	0.1	0
16	7.11.02 1610 - Mav: This open source encyclopedia is an awesome living example of the viral growth of living informatic	1	0	0
17	A new classification table is at Hominid.	1	0	0
18	:: When I'm angry, I can't write from the NPOV. I've let articles I care deeply about languish for over a week at a time,	0.9	0.1	0
19	I'm not sure if it's properly called ''fifth tuning'' or ''perfect fifth tuning'' or what, but it does exist. The octave is slight	1	0	0
20	===Announcement=== I have compared Helga's original version and Helga's condensed version to each other as promis	1	0	0
21	:: You have a valid point,, thanks... any other opinions?	1	0	0

Fig 3: Dataset

B. Training And Testing:

After the dataset has been preprocessed, the abuse dataset that has previously been classified is subjected to the Natural Language Processing algorithm and OpenCV, and the input picture is compared to the trained model file.

Then, a model file is built, the classification result is displayed, and the profile photo is examined with the database using the Swift algorithm to differentiate between fake profile users and real users.

1) NLP Algorithm

The uncertainties in linguistic structure makes it extremely difficult to design software that accurately interprets the exact meaning of either text or speech input. Phonetic spellings, figures of speech, sarcasm, idioms, metaphors, grammatical structures and jargons —these are just a handful of the language and communication irregularities that take humans years to master which the developers must teach natural language-driven software to comprehend precisely from the beginning if those developments are to be effective.

NLP blends cognitive linguistics with statistical, machine learning, and deep learning techniques. These technology solutions, when combined, enable the user to comprehend human language in terms of text or speech data and 'interpret' its exact implications, replete with the speaker's or writer's purpose and attitude.

Phases of NLP

There are five phases of NLP is represented in the Fig below:

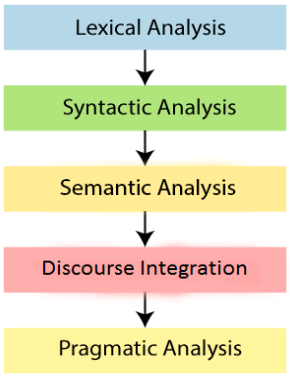


Fig 4: Five phases of NLP

Natural Language Processing APIs enable developers to combine human-machine communications and perform a variety of helpful functions such as speech recognition, chatbots, spelling correction, sentiment analysis, and so on.

2) Swift Algorithm

Due to the fact that they contain difficult-to-read and blunder-prone raw loops, Swift algorithms are effective thinking tools. The setup of the unique Swift algorithm utilized in this detection system is provided in this figure below.

```
each time the search bar text changes {
   create a matcher object with the search text
   create empty array of results

   for each item in the data set to be searched {
      if matcher object indicates the item matches {
         add item to results
      }
   }

   display items in results
}
```

Fig 5: Workflow of Swift algorithm

This algorithm helps to compare the primary user image with those of other accounts in order to spot fraudulent profile usage. As soon as the fake usage is identified, the system bans that specific website. These observations demonstrate the system's ability to swiftly and precisely identify a wide range of faults while working fast and effectively.

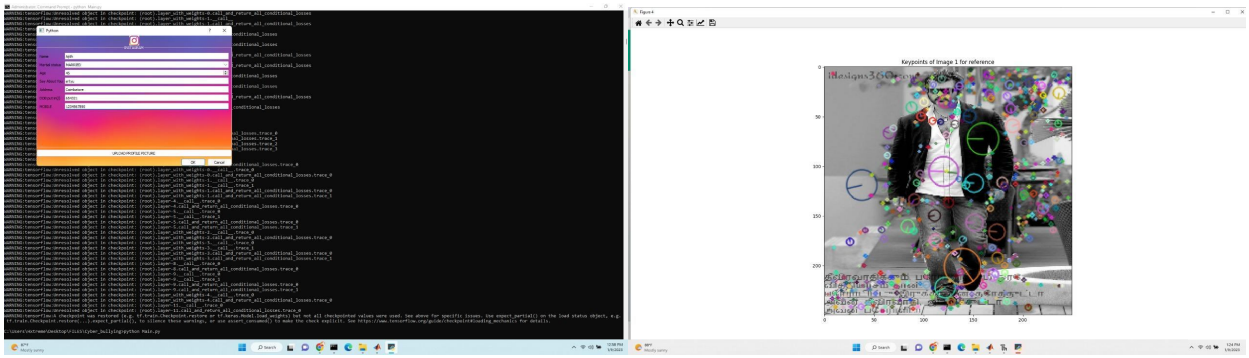


Fig 6: Swift Profile image scanning

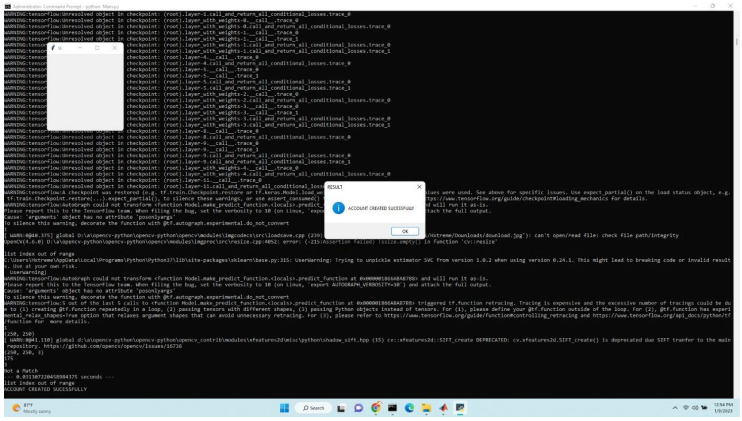


Fig 7: Output

V. FUTURE WORK

Identified, the system bans that specific website. These observations demonstrate the system's ability to swiftly and precisely identify a wide range of In the future, it is hoped to produce a model file, data which will be gathered, and KN machine learning code will be built and trained. The trained model will be used to classify abusive terms, and the Sift algorithm is then used to identify bogus profiles.

VI. CONCLUSIONS

In conclusion, this system enables the model to detect a thorough and organized overview of automatic offensive speech detection, contrasts a few of its existing methods in a systematic way and presents an informative analysis of several published studies on methods for detecting cyberbullying. The volume of hate speech is also rapidly rising as a result of the enormous growth of user-generated web content, notably on social media networks. Here, using the suggested methodology, a trained model file is used in the first phase to detect offensive phrases. The next technique involves recognizing fraudulent profile usage by employing the Swift algorithm to compare the main user picture with other accounts. The system blocks that particular website as soon as the false usage is discovered. These observations show that the system can accurately and quickly identify a diverse array of defects while operating quickly and efficiently.

VII. ACKNOWLEDGMENT

We feel very much grateful to Head of the Department of Computer Science and Engineering, Dr. (Mrs.) S. Sivakumari, School of Engineering, for piloting us properly and stretching out a very big helping hand in the process of accomplishing our project. We also owe our heartfelt thanks to our guide Ms . J.KOWSALYA, Assistant Professor, for stimulating our ideas and opening new horizons for us to learn from our drawbacks in every stage of the project and made us optimistic in carrying our works with her innovative ideas.

REFERENCES

[1] Vaibhav Jain, Ashendra Kumar Saxena, Athithan Senthil, Abhishek Jain, Arpit Jain, “Cyber-Bullying Detection in Social Media Platform using Machine Learning”, 2021, 10th International Conference on System Modeling & Advancement in Research Trends (SMART)

[2] Yeo KhangHsien,ZailanArabee Abdul Salam,VinothiniKasinathan, 'Cyber Bullying Detection using Natural Language Processing (NLP) and Text Analytics', 2022, IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)

[3] J. Yadav, D. Kumar and D. Chauhan, “Cyberbullying Detection using Pre-Trained BERT Model”, 2020, ICESC, pp. 1096-1100, doi:10.1109/ICESC48915.2020.9155700.

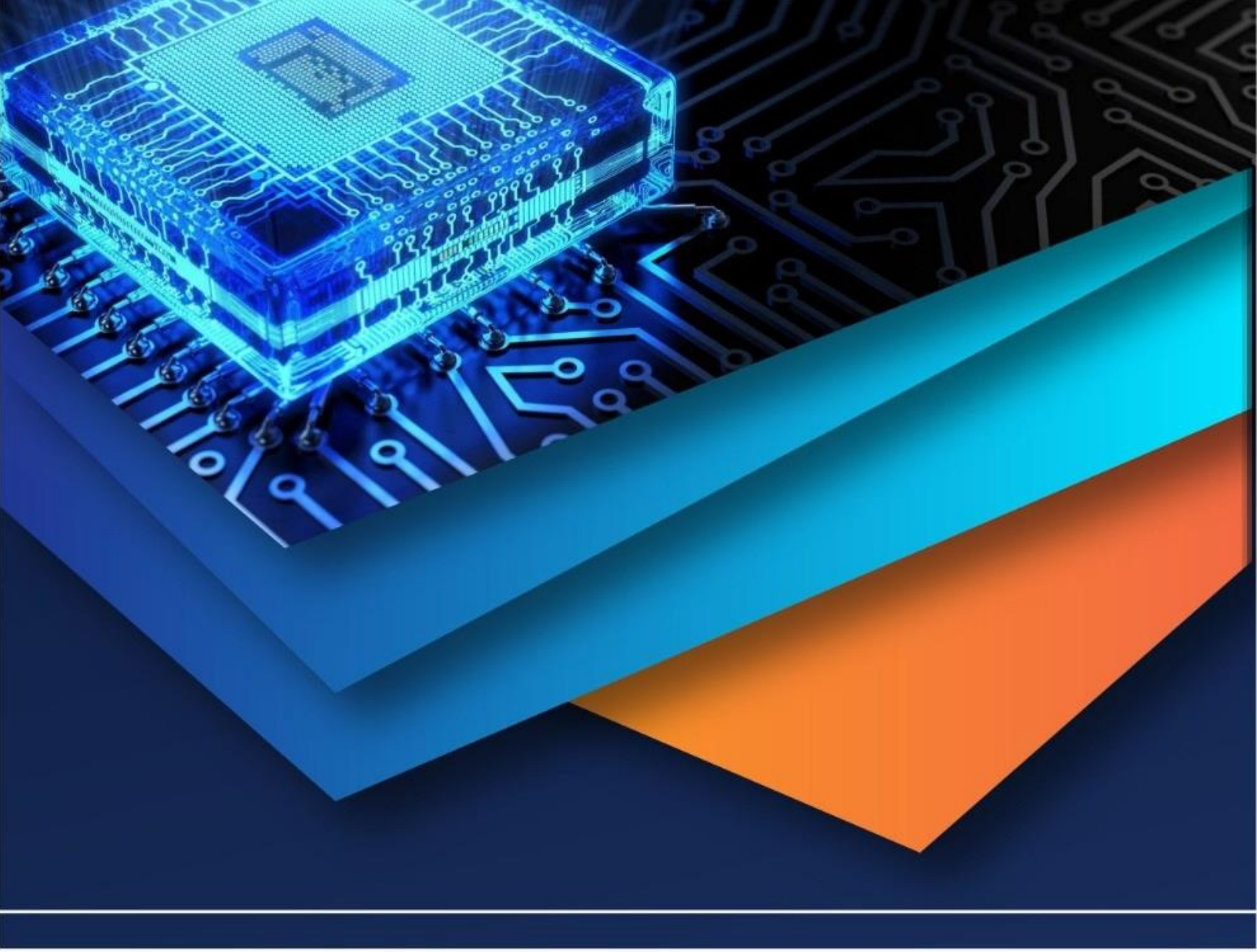
[4] Trana, R.E., Gomez, C.E., & Adler, R., “Fighting Cyberbullying: An Analysis of Algorithms Used to Detect Harassing Text Found on YouTube”, 2020, International Conference on Applied Human Factors and Ergonomics, doi:10.1007/978-3-030-51328-3_2.

[5] T. Vanhove, P. Leroux, T. Wauters and F. De Turck, "Towards the design of a platform for abuse detection in OSNs using multimedial data analysis," 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), Ghent, 2013, pp. 1195-1198

[6] Zhao, Rui et al. “Automatic detection of cyberbullying on social networks based on bullying features.” Proceedings of the 17th International Conference on Distributed Computing and Networking (2016), doi:10.1145/2833312.2849567.

[7] Mitra Behzadi,Ian G. Harris,AliDerakhshan,'Rapid Cyber-bullying detection method using Compact BERT Models',2021 IEEE 15th International Conference on Semantic Computing (ICSC)

[8] ChakcharatNoiklueb, SuraponBoonlue, DaruwanSrikaew, 'Development of Cyber Wellness Assessment Model for Thai elderly population', 2022 International Conference on Digital Government Technology and Innovation (DGTi-CON).
[9] Manne Vinay Kumar,Pulime Satya Sai,A. ThamaraiSelvi, Devarampati Venkata Sai Amarnadh, 'Extracting User Behavioural Control Styles based on Process Mining', 2021 3rd International Conference on Signal Processing and Communication (ICPSC)
[10] Angela AchiaaAikins, Michael Kyobe, 'Towards a conceptual model for developing a mobile application that supports the cyber-arm role of librarians in mitigating mobile bullying', 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)
[11] Hii Lee Jia, Vazeerudeen Abdul Hameed, Muhammad Ehsan Rana, 'CyberSaver â€“ A Machine Learning Approach to Detection of Cyber Bullying', 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)
[12] Amanpreet Singh, Maninder Kaur, "Content-based Cybercrime Detection: A Concise Review", International Journal of Innovative Technology and Exploring Engineering(IJITEE) ISSN: 2278-3075, Volume-8 Issue-8,pages 1193-1207, 2019
[13] R. Zhao, A. Zhou, K. Mao, Automatic detection of cyberbullying on social networks based on bullying features, in Proceedings of the 17th International Conference on Distributed Computing and Networking, 2016.
[14] Zhang, Xiang et al. "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network." 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (2016): 740-745.
[15] Amanpreet Singh, Maninder Kaur, "Content-based Cybercrime Detection: A Concise Review", International Journal of Innovative Technology and Exploring Engineering(IJITEE) ISSN: 2278-3075, Volume-8 Issue-8,pages 1193-1207, 2019
[16] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. "Detecting offensive language in social media to protect adolescent online safety". In Privacy, Security, Risk and Trust (PASSAT),2012 International Conference on and 2012 International Conference on Social Computing(SocialCom), pages 71–80. IEEE, 2012.
[17] K. Jedrzejewski and M. Morzy, "Opinion Mining and Social Networks: A Promising Match," 2011 Int. Conf. Adv. Soc. Networks Anal. Min., pp. 599–604, Jul. 2011.
[18] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network.", Social Informatics - 7th International Conference, SocInfo 2015, Proceedings.
[19] M. Di Capua, E. Di Nardo and A. Petrosino, "Unsupervised cyberbullying detection in social networks", 2016, ICPR, pp. 432-437, doi:10.1109/ICPR.2016.7899672.
[20] Kelly Reynolds, April Kontostathis, LynneEdwards, "Using Machine Learning to Detect Cyberbullying", 2011 10th International Conference on Machine Learning andApplications volume 2, pages 241–244. IEEE,2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY