



Social Media Classifying Toxic Messages Using CNN Text Analysis

Darshan H S¹, Prof. H L Shilpa²

PG Scholar, Dept. of MCA, P.E.S College of Engineering, Mandya, India¹

Assistant Professor, Dept. of MCA, P.E.S College of Engineering, Mandya, India²

Abstract: In an era where a significant portion of interpersonal communication and self-expression transpires online, social media platforms have become rich sources of data on individual psychological states. The research focuses on developing methods to analyse these digital footprints to identify potential signs of mental disorders. Utilizing advanced Natural Language Processing (NLP) techniques and convolutional neural network (CNN) algorithms, we analyse textual content from social networks to detect patterns and markers indicative of mental health issues. We address critical ethical considerations, including user privacy, data security, and the implications of diagnostic accuracy. The findings illustrate the potential of social media mining in providing valuable insights for early mental health intervention. The system distinguishes between toxic and non-toxic words to assess the emotional and psychological well-being of users, thereby enabling more precise and meaningful analysis.

Keywords: Toxic, Non-Toxic comments, Convolutional Neural Network (CNN) text analysis, Natural Language Processing (NLP).

I. INTRODUCTION

Cyberbullying takes various forms, such as circulating rumors on the bases of racism, gender, disability, religion and sexuality, humiliating a person, social exclusion, stalking, threatening someone online and displaying personal information about an individual that was shared in confidence. According to the national advocacy group in US, the bullying can take several forms: racism and sexuality are two of these. Based on a report at Pew Research Centre, two distinct categories of online harassment have been described among internet users. The first category includes less severe experiences: it involves swearing and humiliation, because those who see or experience it often claim they ignore it. The second category of harassment although targeting a smaller number of online users, includes more severe experiences such as physical threats, long-term harassment, trapping and sexual harassment. Assessing the severity level of a cyberbullying incident may be important in depicting the different correlations observed in cyberbullying victims, and principally, how these incidents impact victims' experience with cyberbullying. Hence, many researches try to focus at detecting the cyberbullying automatically for removing those contents from the social media. Though, detecting the cyberbullying in social networking is very challenging task. The system provides a systemic framework for identifying cyberbullying severity in online social networks, which is based on previous research from different disciplines and focuses on a detection technique of exploring on the online social media were many are experiencing the harassments through cyberbully.

This project aims to develop a comprehensive system for detecting mental disorders through social network analysis by leveraging Natural Language Processing (NLP) and machine learning techniques. Social media platforms, while facilitating communication and information exchange, have also become hotspots for cyberbullying and other harmful activities, negatively impacting users' mental health. The system focuses on automatically identifying and filtering damaging or destructive comments using advanced NLP algorithms. By analyzing user-generated content, the system can detect signs of mental disorders such as depression and anxiety. The existing methods primarily concentrate on identifying and classifying cyberbullying content using classifiers like Naïve Bayes and CNNs, often limited by their focus on specific platforms and lack of multilingual support.

Our proposed solution intends to overcome these limitations by employing a diverse, cleaned dataset and training various machine learning models to achieve high accuracy in detection. The model with the best performance is selected to filter and classify toxic comments, ensuring enhanced security and functionality through a 3-Tier Architecture, which restricts unauthorized access to sensitive data. The application is developed using Python and ASP.NET, with tools like Microsoft Visual Studio and SQLyog, ensuring a robust and scalable solution to mitigate the adverse effects of harmful social media interactions and support mental health monitoring.



II. RELATED WORK

B. Sri Nandhini and et al [1] proposed a framework for observing and grouping cyberbullying activities such as provocation, blazing, bigotry and psychological oppression. The algorithm utilized in this examination was Naive Bayes classifier for arranging the cyberbullying movement as well as Lowenstein calculation for cyberbullying identification. Naïve Bayes classifier used for cyberbullying activities & the mean exactness gotten from conspiring [1].

Zahra Ashktora b and et al [2] focuses on tackling and addressing ways to mitigate depression, suicide and anxiety resulting from the occurrence of cyberbullying. Effectiveness in using technological procedures and mechanisms to curb cyberbullying with the aid of tertiary prevention on the Instagram social media platform. Evaluating the effectiveness of cyberbullying mitigation techniques on Instagram.

Online Social Network (OSN) services, such as Facebook, Twitter, and MySpace are gaining in popularity as a main source of spreading messages to other people. Messaging is widely used and very useful in various purposes, for example, business, education, and socialization. However, it also provides opportunity to create harmful activities. There are numerous evidences showing that messaging can introduce the very concerned problem, namely cyberbullying [3].

Gamming chat is presented that continuously in-game chat data from one of the most popular online multi-player games: World of Tanks. The data was collected and combined with other information about the players from available online data services. It presents a scoring scheme to enable identification of based on current research. Classification of the collected data was carried out using simple feature detection with SQL database queries [4].

Social networking platforms are being widely used These days for more than one functions like leisure, Networking, and many others., and turning into a boon for each person but on Any other side, with the increasing range of customers on social media leads to a new way of cybercrimes. Cyberbullying Is turning into a main problem and is defined as an intentional or a competitive act that is completed by using someone or Groups of people using repeated communication Additional time against a victim who cannot effortlessly shield him or Herself. With the inception of the internet, it become only a Rely of time until bullies determined their manner onto this new and opportunistic platform [5].

Social networking sites indications to the problematic habit. It offers an opportunity to identify disorder at an early stage. These MDD system are made a different and advanced for the preparation of disorder detection. It is difficult to detect disorder because the mental state cannot be observed directly from the registers of online social activities. Social Network Mental Disorder (SNMD) users can be automatically identified and classified into various categories like Virtual Relationship Addiction, Obsessive Online Gambling and Information Glut using SNMD based tensor model, with the data sets collected from data logs of various Online Social Networks (OSNs) [6].

Social networks (SNs) by exchanging information, delivering comments, finding new information, and engaging in discussions. These data, available in various forms, such as images, text, and videos, may be interpreted to reflect the user's activities, including their mental state regarding depression. Depression is a chronic disease from which the vast majority of users suffer, and it has emerged as a significant issue relating to mental health on a global scale. Even though several procedures have been utilized over the past few decades to diagnose depression, machine learning (ML) and deep learning (DL) techniques supply superior insights [7].

Healthcare databases now contain significantly more data sources in terms of volume and function. In the field of health informatics, analyzing this massive amount of medical data presents both opportunities and challenges for knowledge creation. Over the past ten years, scientific domains such as healthcare and medicine have increasingly employed social network analysis methods and community discovery algorithms. While community detection algorithms are frequently employed in social network research, there is still a dearth of thorough reviews of their applications in healthcare that would help the field of health informatics as well as medical professionals [8].

Alireza Pourkeyvan and et al [9] diagnosis of mental disorders and intervention can facilitate the prevention of severe in their study uses social media and pre-trained language models to explore how user-generated data can predict mental disorder symptoms. BERT models of Hugging Face with standard machine learning techniques used in automatic depression diagnosis in recent literature.

Sentiment analysis is to discover the exactness of the underlying emotion in a given situation. It has been applied to various Felds, including stock market predictions, social media data on product evaluations, psychology, the judiciary, forecasting, illness prediction, agriculture, and more.



Many researchers have worked on these topics and generated important insights. These outcomes are useful in the fields because they (outcomes) help people comprehend the general summary quickly. Additionally, sentiment analysis aids in limiting the harmful effects of some posts on various social media sites such as Facebook and Twitter [10].

The literature survey presents various approaches and techniques for detecting and addressing cyberbullying and related mental health issues. One framework utilizes a Naive Bayes classifier and the Lowenstein algorithm, achieving high precision in identifying cyberbullying activities on platforms like Formspringme and Myspace. Another study focuses on mitigating depression, suicide, and anxiety from cyberbullying on Instagram using Naive Bayes and tertiary prevention techniques. A third work uses K-means clustering and Naive Bayes for classifying polite and abusive messages on Twitter. Additionally, an automatic system collects in-game chat data from World of Tanks to identify cyberbullying through AI-based sentiment analysis and SQL queries. Another approach involves preprocessing text data for cyberbullying detection using machine learning. Research on mental disorder detection in social networks proposes a machine learning system to identify disorders early. A systematic review highlights the use of ML and DL techniques for depression detection. Social network analysis and community detection algorithms are reviewed for their healthcare applications.

III. PROPOSED SYSTEM

The proposed system to deal with cyberbullying includes: Damaging or destructive writings, comments on posts of individuals are detected and filtered out using Natural Language Processing and Machine learning algorithms. We intend to use a diverse dataset, clean it and feed it to different machine learning algorithm models to obtain the accuracy of each algorithm. The algorithm with high score is selected and implemented to detect and classify the cyberbullying keywords. Each comment posted is fed into the selected algorithm and evaluated to mask the toxic comments from viewers.

The image outlines a system architecture (figure 1) for detecting mental disorders via online social media mining. At its core is a social network application where users can register, log in using their email or mobile number, and upload images. Users can view their own and others uploaded images, as well as post and view comments on these images. The system utilizes Python for CNN (Convolutional Neural Network) text analysis to process and analyze the data. The analyzed data and user interactions are stored in a MySQL database, which supports the backend operations of the application. This architecture aims to identify mental health disorders by analyzing User-generated content on the social network.

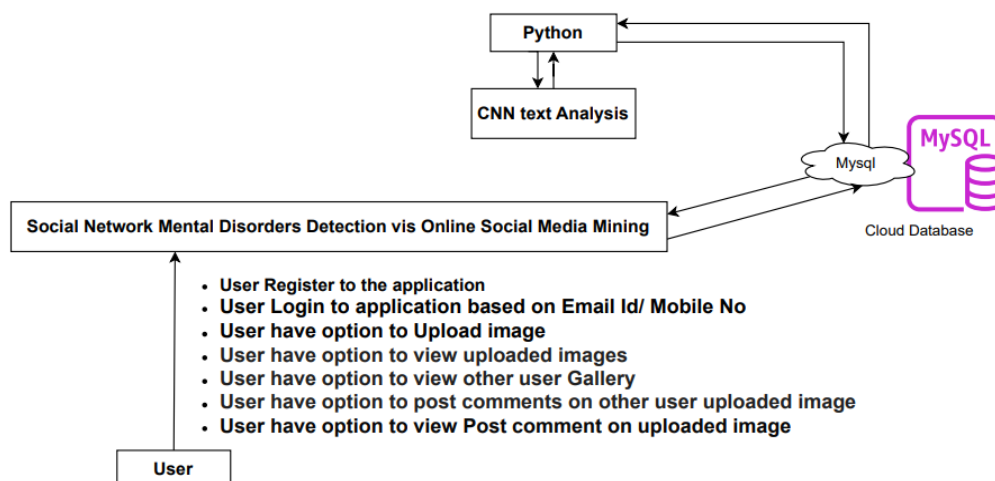


Figure 1: System architecture

NATURAL LANGUAGE PROCESSING (NLP)

The collected textual data undergoes NLP processing. This involves several key tasks, such as

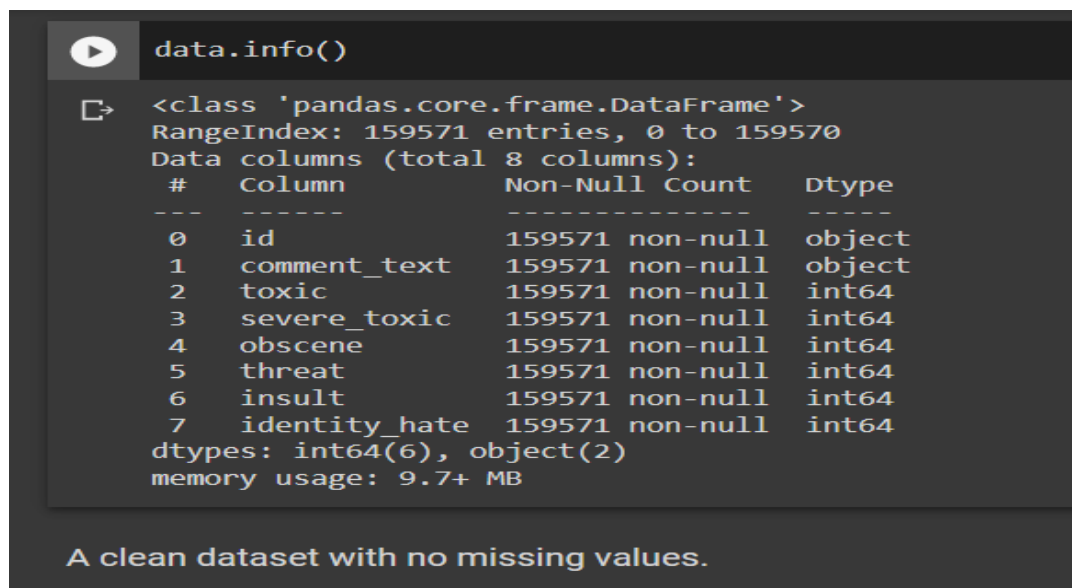
- **Tokenization:** Breaking down the text into individual words or tokens.
- **Stop Words Removal:** Removing common but insignificant words (e.g., “and”, “the”) that do not contribute to the meaning.



- **Normalization:** Converting text to a standard format (e.g., lowercasing, stemming, or lemmatization).
- **Sentiment Analysis:** Analysing the emotional tone of the text (e.g., positive, negative, or neutral).
- **Feature Extraction:** Identifying and extracting relevant features or patterns that can be used for further analysis by the CNN model.

DATA PRE-PROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. Data preprocessing is to handle missing data in the datasets. If the dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.



```
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 159571 entries, 0 to 159570
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  -
 0   id              159571 non-null  object
 1   comment_text    159571 non-null  object
 2   toxic           159571 non-null  int64
 3   severe_toxic    159571 non-null  int64
 4   obscene         159571 non-null  int64
 5   threat          159571 non-null  int64
 6   insult          159571 non-null  int64
 7   identity_hate   159571 non-null  int64
dtypes: int64(6), object(2)
memory usage: 9.7+ MB

A clean dataset with no missing values.
```

Figure 2: Data Pre-processing

TEXT-PREPROCESSING

In text preprocessing step the stop words, numbers, all non-ascii characters, punctuation are removed using Natural Language Toolkit (NLTK) library.

```
# Text preprocessing steps - remove numbers, capital letters, punctuation, '\n'
import re
import string

# remove all numbers with letters attached to them
alphanumeric = lambda x: re.sub('\w*\d\w*', ' ', x)

# '[%s]' % re.escape(string.punctuation), ' ' - replace punctuation with white space
# .lower() - convert all strings to lowercase
punc_lower = lambda x: re.sub('[%s]' % re.escape(string.punctuation), ' ', x.lower())

# Remove all '\n' in the string and replace it with a space
remove_n = lambda x: re.sub("\n", " ", x)

# Remove all non-ascii characters
remove_non_ascii = lambda x: re.sub(r'^\x00-\x7f', r' ', x)

# Apply all the lambda functions wrote previously through .map on the comments column
data['comment_text'] = data['comment_text'].map(alphanumeric).map(punc_lower).map(remove_n).map(remove_non_ascii)

data['comment_text'][0]
```

Figure 3: Text-Preprocessing



IV. RESULT AND DISCUSSION

```

Anaconda Prompt - python c X + v
(base) C:\Users\abhij>cd\toxic
(base) C:\toxic>python cnn.py
Connected successfully
Good pic
{'text': 'Good pic', 'class': 'Non-Toxic'}

Anaconda Prompt - python c X + v
(base) C:\Users\abhij>cd\toxic
(base) C:\toxic>python cnn.py
Connected successfully
Bad Pic
{'text': 'Bad Pic', 'class': 'Toxic'}
Bad Pic updated to post

```

Figure 4: Result Analysis for classifying the comments

The terminal likely belongs to a system that classifies or processes text input. The specific input “Good pic” was analysed and categorized as “Non-Toxic”, meaning it is considered harmless or appropriate in this context. This kind of classification could be part of a content moderation or text classification system designed to filter out toxic or inappropriate content.

The process where a text input is analyzed by a Python script, and based on the content, it is categorized as either “Toxic” or not. In this case, the input “Bad Pic” was classified as “Toxic”, and the system took a specific action in response, such as not updating or flagging a post on user Interface (UI). This is likely part of a content moderation or filtering system aimed at identifying and managing toxic content in an online environment.

V. CONCLUSION

The exploration of social network mental disorders detection through online social media mining opens a new frontier in the field of mental health care. This innovative approach harnesses the vast and growing expanse of digital footprints left on social media platforms, providing a unique opportunity to identify and analyze signs of mental health issues.

The methodology, which integrates advanced data mining techniques, natural language processing, and Convolutional Neural Network algorithms, offers a promising avenue for early detection, large-scale monitoring, and potentially more personalized mental health care. The proposed system can be improved by adding context based cyberbully detection that is system detecting sarcasm, and sentiment in comments. The system can also improve by detecting multiple language cyberbully contents.

REFERENCES

- [1]. B. Sri Nandhini, J.I. Sheeba. 2015. Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45, pp.485-492.
- [2]. Zahra Ashkora b. “A Study of Cyberbullying Detection and Mitigation on Instagram” Companion: Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion.
- [3]. Walisa Romsaiyud, Kodchakornna Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, Pirom Konglerd. “Automated cyberbullying detection using clustering appearance patterns”. 'Summary of Our Cyberbullying Research, Available: <http://cyberbullying.org/summary-of-our-cyberbullyingresearch>.



- [4]. Shane Murnion, William J. Buchanan, Adrian Smales, Gordon Russell. "Machine learning and semantic analysis of in-game chat for cyber bullying". Automatic analysis and identification of verbal aggression and abusive behaviors for online social games.
- [5]. John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer. "Social Media Cyberbullying Datacom using Machine Learning". International Journal of Advanced Computer Science and Applications.
- [6]. Chandrakant Mallick, Sarojananda Mishra, Parimal Kumar Giri and Bijay Kumar Paikaray. "Machine learning approaches to sentiment analysis in online social networks". International Journal of Work Innovation Vol. 3, No. 4.
- [7]. Ferdaous Benrouba, Rachid Boudour. "Emotional sentiment analysis of social media content for mental health safety".
- [8]. Ringsquandl M, Petkovic D (2013) Analyzing political sentiment on Twitter. In: Proceedings of the 2013 AAAI spring symposium series, Stanford, pp 25–27.
- [9]. Alireza Pourkeyvan; Ramin Safa; Ali Sorourkhah. "Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks".
- [10]. Mehrdad Rostami; Mourad Oussalah; Kamal Berahmand; Vahid Farrahi. "Community Detection Algorithms in Healthcare Applications: A Systematic Review".