Identifying Cyberbullying and Hate Speech via Sentiment Analysis of Social Media Text Social Networks

Loveleen Kaur Pabla[1], Prashant Kumar Jain[2], Prabhat Patel[2] and Shailja Shukla[3]
[1]*Department of Information Technology, Jabalpur engineering College, Jabalpur, Madhya Pradesh, India*
[2]*Department of Electronics and Telecommunication, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, Madhya Pradesh, India*
[3]*Professor & Head, Department of Electrical Engineering, Jabalpur engineering College Jabalpur, Madhya Pradesh, India*

Keywords: Sentiment Analysis, Text Mining, Natural Language Processing, Cyber Bulling, Hate Speech.

Abstract: In the current digital era, in addition to being a vital component of modern life, social media has also given rise to the pervasive and dangerous issue of cyberbullying. The rise of social media platforms has coincided with an increase in cyberbullying, a type of abuse that takes place on these sites. Social networks on the internet are expanding quickly in many ways. These platforms are being utilised by various businesses to advertise their goods and interact with final customers. Conversely, some groups are abusing these platforms to spread hate speech, violent materials, and offensive or harassing material. The social climate both online and offline is out of balance as a result of all these events. The proposed work analyses the text content of social media posts, messages, and blogs to provide a research proposal on the identification of hate and bullying content online. Machine learning methods and sentiment analysis are used in this context as crucial instruments to handle large volumes of data. In addition to finding bullying terms, sentiment analysis has the ability to identify victims in danger of hurting themselves or others. In order to detect signs of bullying in social media posts, our research relies on applying deep learning and natural language comprehension approaches. A Recurrent Neural Network with Long Short-Term Memory (LSTM) cells was constructed using multiple embeddings. One method makes use of BERT embeddings, while the other uses OpenAI's freshly published embeddings API in place of the embeddings layer.

1 INTRODUCTION

Hate speech and cyberbullying continue to occur in a similar way to traditional forms. Direct bullying or hate speech simply altered their appearance and had a detrimental effect on the victim's mental state. This social threat is becoming more widespread every day on a variety of social media sites, not just in India but globally as well. Teenagers are not the only age group affected by these things; they work for all ages. It is unacceptable to disseminate the dubious text by various means of communication, such as social media, SMS, emails, and others. Thus, text analysis methods are helpful for examining texts and identifying patterns of bullying and hatred. These days, these methods are used for prediction, recommendation, and decision making in a variety of application domains, including business, banking, education, engineering, and medicine. However, assessing the attitudes in a document requires more than just basic text analysis and categorization methods. As a result, NLP-based text analysis is employed to identify the attitudes and emotions contained within the text. Understanding text sentiments requires a solid understanding of natural language processing (NLP). Human life preservation may benefit from it. Finding text communication's sentiment classes and intensity scores is the main goal of the suggested effort. Finding the subclasses of sentiment classes is the goal of the technique, which offers more than just sentiment classification. provides a sentiment score as well, which aids in determining the degree of hate. This work focusses on the domains of natural language processing (NLP) and machine learning. NLP is a field that deals with in-text blocks, text classification, and emotion detection. Sentiment analysis can be used to content found on forums, microblogs, e-commerce product

Pabla, L. K., Jain, P. K., Patel, P. and Shukla, S.
Identifying Cyberbullying and Hate Speech via Sentiment Analysis of Social Media Text Social Networks.
DOI: 10.5220/0013899800004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 3*, pages 448-453
ISBN: 978-989-758-777-1

reviews, social media, and other sources. This effort analyses text from a variety of public sources using machine learning methods. Therefore, a suggestion to build and implement a hate speech and cyberbullying detection system is put forth. There have been numerous contributions in this area recently from different authors. However, the majority of them can only forecast positive or negative orientations and their subclasses. However, there are no appropriate methods to address the influence of emotions. Therefore, the suggested task entails creating a data model to gauge how strongly the bully or hate speech is conveyed in the text. We also attempted to subdivide the emotion classes based on the feelings' intensity. Social media has become an essential component of everyday life in recent years, enabling previously unheard-of levels of connection and communication. But it has also played a part in the increase in online harassment and cyberbullying (Feinberg and Robey, 2009). Cyberbullying, in contrast to traditional bullying, can occur at any time, and because of the anonymity offered by the internet, offenders are encouraged to behave without fear of instant repercussions (Hasan et al., 2023). When compared to in-person bullying, this anonymity frequently results in higher rates of cyberbullying.

2 NATURAL LANGUAGE PROCESSING

There have been numerous contributions in this area recently from different authors. However, the majority of them can only forecast positive or negative orientations and their subclasses. However, there are no appropriate methods to address the influence of emotions. Therefore, the suggested task entails creating a data model to gauge how strongly the bully or hate speech is conveyed in the text. We also attempted to subdivide the emotion classes based on the feelings' intensity. Natural language describes the speech and text that humans use to communicate with one another. Text is all around us, and we use menus, signage, emails, SMS, websites, and more. We communicate via writing and speaking to one another. Speaking might be simpler to learn than writing. We connect with one another via text and voice. Like other forms of data, natural language requires tools for comprehension and reasoning.

- **Automatic Summarization**- Data is the most valuable resource in this day and age. We obtain the necessary and helpful quantity of information from a variety of sources. We have access to far more knowledge than we can comprehend, and the amount of information is overwhelming. Since there will always be an abundance of information available online, we require an automatic text summarization system. The process of producing a concise, precise synopsis of text documents is known as text summarization. In less time, text summarization will provide us with pertinent information. Automatic text summarization relies heavily on natural language processing.
- **Question-answering**- Question-answering (QA) is another use case for natural language processing. Although search engines make the world's knowledge accessible, they nevertheless fall short when it comes to providing answers to our queries. This is the route that businesses like Google are taking. It is a branch of NLP and AI. Its main goal is to create systems that can automatically respond to human enquiries. In order to create valid answers, a computer system that comprehends the NL can transform the sentences into an internal representation. Syntax and semantic analysis can produce the precise answers. Among the difficulties NLP has in creating an effective QA system are lexical gaps, ambiguity, and multilingualism.
- **Sentiment Analysis**- Sentiment analysis is another way that NLP is used. To determine the sentiments of several posts, sentiment analysis is utilised. When feelings are not clearly conveyed, it is used to determine the sentiment. Sentiment analysis is an NLP application that businesses use to find out what their customers think. Knowing what their clients think of the goods and services will be beneficial. They can use sentiment analysis to assess their reputation based on consumer reviews. Sentiment analysis helps us better understand the motivations behind the opinions expressed by analysing sentiments in context.

3 HATE SPEECH AND CYBER BULLING

Targeting an individual or a group on the basis of their religion, race, ethnic origin, handicap, sexual orientation, or gender is known as online hate speech, and it typically occurs on social media. It is where several tensions converge: [weasel words]. It is vivid when diverse groups express themselves differently from one another. Multilateral and multi-stakeholder procedures is a broad word. Hate speech is becoming a common term and is growing daily. Threats to the security of both individuals and groups are being mixed together. Google, Twitter, Facebook, and other social media sites started defining hate speech in their own ways. It is necessary for national and regional bodies to raise awareness of such phrase. Because of the speed and breadth of the Internet, governments find it difficult to enforce laws in this situation. Because of the problems with hate speech, private channels for expression that operate in public spaces, like Facebook and Twitter, have become more popular. What is hate speech, how it relates to offline and success, what constitutes apparent hate speech, and likely punishment have led to a lot of discussion about the solution and how it should be based. However, this emphasis has also hindered more thorough efforts to identify the root reasons of the phenomena.

Bullying or harassment via digital devices, such as smartphones, laptops, and computers, is known as cyberbullying. Chat rooms, social networking, and gaming are examples of platforms where cyberbullying can happen since users can read and share content. The various forms of bullying include abusive remarks, text messages, and messages. It includes sending, posting, and disseminating inaccurate or unfavourable information. Do you think your child has a smartphone addiction? Has he or she grown more reclusive and socially insensitive? Changes in a child's behaviour might cause parents great concern. Watch out! It's possible that your youngster is being bullied online. Parents can report cyberbullying according to research on the topic. Some typical forms of cyberbullying are as follows:

- Posting offensive or derogatory remarks
- rumours, or pictures about someone
- making a phoney or offensive webpage about someone
- threatening or inciting someone online
- inciting hate speech about someone based on their race, religion, ethnicity, or politics
- posing as someone else online in order to post personal or false information about someone

Information about someone's social media accounts is obtained and utilised for online abuse, slander, and other purposes. Cyberbullying has seen a sharp increase in India as social media and reasonably priced broadband services have become more widely available. Studies show that over eight out of ten people experience some form of cyberbullying. Of these, around 63% were subjected to online harassment and abuse, and 59% were the target of unfounded rumours and gossip. Furthermore, a number of research show that India has the greatest rate of cyberbullying in Asia. In particular, 50% of Indian city women experience abuse online.

4 PROPOSED WORK AND PROPOSED DATA MODEL

The proposed work's goal is to research and enhance the sentiment analysis method for identifying hate speech and posts related to cyberbullying on social media. Anyone can post anything on social media. Therefore, the following goals are set in order to keep an eye on and maintain the cleanliness of social networks.

- **To investigate and learn about hate speech and cyberbullying detection methods**- A survey of the current methods and resources for sentiment analysis, hate speech, and cyberbullying is necessary because there has been a lot of work done recently on social media text processing for sentiment label recovery. As a result, this survey aids in the formulation of our solution development plan.
- **For automatic sentiment identification, to expand and enhance current sentiment analysis methods**- The promising classification methods are retrieved from the literature review in order to improve their performance in sentiment classification. Furthermore, the updated features allow for precise detection of hate speech and bullying.
- **Incorporating sentiment's intensity and semantics into tex**t- Here, the sentiment analysis method that was previously provided is changed to determine the sentiment impact on the target individual or

community. As a result, a scale and score are created to quantify the hate and bad attitudes expressed in the text.

- **To research and evaluate the suggested cyberhate detection framework's performance in comparison to cutting-edge methods-** The rationale behind the suggested model is attempted to be illustrated. Furthermore, a performance comparison between the suggested work and pertinent existing methodologies is shown. Additionally, utilising the information and facts that were available on Twitter, we attempted to follow an actual case.

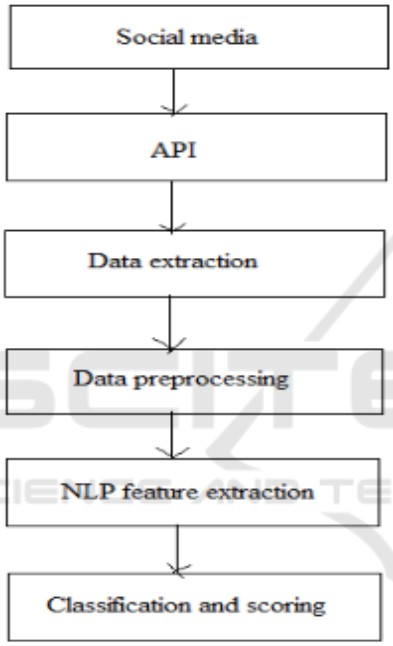

Figure 1: Proposed Data Model.

Figure 1 shows the proposed data model. The goal of the proposed work is to enhance the sentiment analysis method already in use. A score system based on delivered text will be incorporated into the extension to gauge the sentiment intensity. Cyberbullying and hate speech content detection applications are utilised in conjunction with this scoring system. Thus, image 1 illustrates a basic model. Twitter is therefore chosen as the social media platform for data collection and experimentation. In this context, we can locate a variety of offline and online dataset resources. Furthermore, social media sites like Facebook and Twitter offer web APIs for retrieving data from live feeds. We must enhance the quality of the raw offloaded data after offloading the necessary data. Thus, in order to improve the quality of the data, pre-processing is used first. Since the offloaded data is essentially in text format, or unstructured format, it comprises a variety of undesirable characters and stop words. The offline contents were attempted to be removed by the pre-process. Following content refining, feature extraction methods are applied. Here, we must identify a few text-based or natural language processing (NLP) feature selection techniques and assess their efficacy before implementing them in the feature extraction of the suggested model. To identify the desired patterns, supervised, semi-supervised, and unsupervised learning approaches are used with the computed features from the pre-processed text data. This procedure is referred to as the suggested system's training. Lastly, the sentiment class for the provided data is generated by the suggested model. In essence, we are unable to judge a social media post's criticality based on its polarity, or whether it is accurate or not. In order to determine the severity and negativity of a hate speech post, we are attempting to calculate a sensitivity score utilizing the various facts that are currently accessible. An overview of the suggested data model is presented in this paper, and the specifics of the model are provided in the next piece.

5 EXPERIMENTATION

Algorithm: Sentiment Analysis using BERT or OpenAI Embeddings in RNN

1. Input: Input text T = {t1, t2, . . . , tn}, where ti represents each token in the input sentence.
2. Step 1: Tokenization

Tokenize the input text T using:

- BERT: Use BERT's tokenizer to obtain subword tokens Tsub = {s1, s2, . . . , sm}.
- OpenAI: Input the raw text T directly to OpenAI's embedding API.

3. Step 2: Embedding Extraction

Extract contextual embeddings using:

- BERT: Pass tokenized text Tsub through a pre-trained BERT model to obtain embeddings E = {e1, e2, . . . , em}.
- OpenAI: Use OpenAI's API to get high-dimensional vector embeddings E = Embedding.create(input = T).

4. Step 3: Classification Model

Input the embeddings E into a RNN model to perform the classification task.

5. Step 4: Output

The final output is the predicted class y_{pred} with the highest probability.

Binary Cross Entropy Loss (BCELoss), which works well for binary classification tasks, is used to train the model. The definition of the BCELoss is

$$L = (y \log(p) + (1 - y) \log(1 - p)) \qquad (1)$$

Where, the anticipated probability for the positive class is denoted by p and the true label by y. The classifier performs better when BERT-base or OpenAI embeddings that is our hybrid architectures are used than when the RNN is used alone.

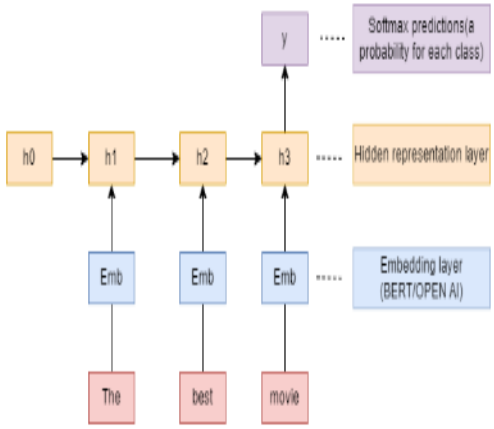

Figure 2: Bert/Openai Api Embeddings in an Rnn Network.

Providing high-quality feature vectors that were taken from the text data allows for this improvement. The RNN network with BERT or OpenAI embeddings is displayed in Figure 2.

Implementation Details- Training and testing data were divided in an 80:20 ratio. We employed dropout to reduce overfitting and gradient clipping to avoid exploding gradients, with a learning rate of 0.001. The training procedure, which lasted ten epochs, used the Adam optimiser.

6 CONCLUSIONS

Bullying and hate are becoming more and more of a social threat worldwide, manifesting in different ways such as race, religion, body, physical ability, and others. It is overly delicate and has a detrimental effect on a person's thinking as well as the target community. Furthermore, hatemongers are using digital communication platforms and social media to spread their malicious content. In this regard, the suggested work is an endeavour that aids in the collection and examination of textual content from widely used social media platforms in order to identify contents that are counterfeit. The suggested effort first identifies the questionable content, preventing it from going viral across many channels and ultimately saving many lives.

- To accomplish the necessary goal, the following next stages are included in the suggested sentiment analysis based on bullying and hate speech detection technique:
- To identify appropriate sentiment-based feature selection methods and text feature selection methods, do a comparative performance analysis.
- To handle the text input and pertinent feature set, find a classifier that is both accurate and efficient.
- Look for some engineered qualities that contribute to the legitimacy of the source.For a target event, identify the origin of compromised content and potential distribution channels.

Our research showed that in capturing the nuances of cyberbullying language, both OpenAI and BERT embeddings performed noticeably better than a simple RNN model. OpenAI embeddings were the most effective of them all, demonstrating their exceptional capacity to comprehend sentiment and context in user-generated information. Going forward, the dataset's modest size presents difficulties for model performance. In order to solve this, we intend to investigate zero-shot or few-shot learning techniques in subsequent research, which may improve the model's generalisation and precision in forecasting cases of cyberbullying with a small amount of labelled data.

REFERENCES

A. kumar, "What is data mining", CSE Dept, Dr. APJ Abdul Kalam UIT Jhabua (M.P.) "Investment: Unit – 1, Security Analysis and Portfolio Management", http://www.pondiuni.edu.in/storage/dde/downloads/ finiv_ sapm.pdf

B. Krawczyk, "Learning from imbalanced data: open challenges and future directions", Prog Artif Intell (2016) 5:221–232

C. Clavel, Z. Callejas, "Sentiment Analysis: From Opinion Mining to Human-Agent Interaction", IEEE Transactions on Affective Computing, VOL. 6, NO. X, XXXXX 2015 "Preparing for the Future of Artificial Intelligence", Executive Office of the President National Science and Technology Council Committee on Technology, Oct. 2016, https:// obamawhitehouse. archives.gov/sites/default/files/whitehouse_files/micro sites/ostp/NSTC/preparing_ for_the_future_of_ai.pdf

D Cross. 2008. Cyberbullying: International comparisons, implications, and recommendations. In 20th biennial meeting of the International Society for the Study of Behavioural Development, Wurzburg, Germany.

G. W. Blood, I. M. Blood, "Cyberbullying: Responsibility, Concerns and Personal Experiences of School-based Speech-Language Pathologists", International Journal for Infonomics (IJI), Volume 9, Issue 1, March 2016

G. Peters, R. Weber, “DCC: a framework for dynamic granular clustering”, Granul. Comput. (2016) 1:1–11, DOI 10.1007/s41066-015-0012-z

I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N. R. Ke,S. Rajeshwar, A. d. Brebisson, J. M. R. Sotelo, D. Suhubdy, V. Michalski, A. Nguyen, J. Pineau, Y. Bengio, “A Deep Reinforcement Learning Chatbot”, arXiv:1709.02349v2 [cs.CL] 5 Nov 2017

J. Wan, D. Wang, S. C.H. Hoi, P. Wu, J. Zhu, Y. Zhang, J. Li, “Deep Learning for Content-Based Image Retrieval: A Comprehensive Study”, Proceedings of the ACM International Conference on Multimedia: November 3-7, 2014, Orlando. pp. 157-166

J. Jayasree, Sri M. A. Mathi, S. Malini, E. Bavithra, Dr. P. Boobalan, “Product Ranking System in E-Commerce Website for Validation using Sentimental Analysis”, International Journal of Science Technology & Engineering, Volume 3, Issue 09, March 2017

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics.

Jalal Omer Atoum. 2020. Cyberbullying detection through sentiment analysis. 2020 International Conference on Computational Science and Computational Intelligence (CSCI), pages 292–297.

Jalal Omer Atoum. 2021. Cyberbullying detection neural networks using sentiment analysis. 2021 International Conference on Computational Science and Computational Intelligence (CSCI), pages 158–164.

M. Rajan, T. S. Rinku, V. Bhojane, “Information Retrieval in Malayalam Using Natural Language Processing”, International Journal of Scientific & Engineering Research, Volume 5, Issue 6, June-2014

M. M. Kampert, J. J. Meulman, J. H. Friedman, “rCOSA: A Software Package for Clustering Objects on Subsets of Attributes”, Journal of Classification 34:514-547 (2017)

M. M. Mironczuk, J. Protasiewicz, “A recent overview of the state-of-the-art elements of text classification” ,Expert Systems With Applications, 106, 2018, 36-54

Michael Agbaje and Oreoluwa Afolabi. 2024. Neural network-based cyber-bullying and cyber-aggression detection using twitter(x) text. Revue d’Intelligence Artificielle.

Nhan Cach Dang, María N. Moreno García, and Fernando de la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. ArXiv,abs/2006.03541.

S. Zhang, C. Zhang, Q. Yang, “Data Preparation For Data Mining”, Applied Artificial Intelligence, 17:375– 381, 2003, Copyright # 2003 Taylor & Francis, 0883-9514/03$12.00+.00,DOI:10.1080/0883951039021926 “DataMiningOverview”,https://www.tutorialspoint.com/data_mining/dm_overvie w.htm

S. Khemka, “Why Data Mining? Editor’s Point of View”, digital valley, Vol.1, No 4, July 2018

Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec, volume 10, pages 2200–2204. Valletta.

T. Wuest, D. Weimer, C. Irgens, K. D. Thoben, “Machine learning in manufacturing: advantages, challenges, and applications”, Production & Manufacturing Research: An Open Access Journal, 2016, VOL. 4, NO. 1, 23–45

U. Fayyad, G. P. Shapiro, P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine Volume 17 Number 3 (1996) (© AAAI)

V. Gupta, G. S. Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, Aug 2009

W. Medhat, A. Hassan, H. Korashy, “Sentiment analysis algorithms and applications: A survey”, Ain Shams Engineering Journal 2014, 5, 1093-1113

Y. Roh, G. Heo, S. E. Whang, “A Survey on Data Collection for Machine Learning”,arXiv:1811.03402v2 [cs.LG] 12 Aug 2019