Leveraging Deep Neural Networks for Real-Time Detection of Cyberbullying and Offensive Memes in Social Network

K. Jayasurya, T. Periyasamy, S. Prasanth, S. Sibhisaran and A. Ananya
Department of Artificial Intelligence and Machine Learning, M. Kumarasamy College of Engineering, Karur, Tamil Nadu, India

Keywords: Social Media, Freedom of Speech, Hate Speech, Cyberbullying, Sentiment Analysis.

Abstract: Although connection to other people is possible through the use of social media, the result has been what the social media users use to call as freedom of speech, which means that hate speech is also possible, and cyber bullying. Free speech has come to mean liberty to post vengeful memes or to harass individuals or groups based on their gender, colour, or religion. To locate such risky content, this study recommends a two-part approach that involves natural language processing and optical character recognition. The model takes textual and the visual input through deep neural recognition and analyses the sentiment of the text through a sentiment recognition dictionary and a sentiment estimator. Negative content is filtered according to script and feedback provided by the users of the application. When implemented on social media networks, these insights, which are delivered in real-time, can significantly improve the fight against the use of technology to perpetrate abuse.

1 INTRODUCTION

But most important is – how open and available the SM communication is. Open one or several accounts in social networks and begin to share with other people opinions, ideas and beliefs. But it has now become a crucial facet of the current society because modern society does not know it. As they said that all those people who spread hate in one way or the other, for example in politics, racists etc. are the ones who are most likely to become either abusive or harassing towards them in such a type of stage. Some ideas like the use of politics or racism usually occur via the various social media platforms. Unfortunately, such purposes are increasingly often used for such actions as blackmail, abuse, and even computer criminality. By this it has become easier to know and interact with other communities or organizations of like nature through networks. Slightly above a third of these network users are below 30 years of age. These platforms are filled with a wealth of information to enable researchers undertake comprehensive analysis within targeted sub-topics of study. The process of identifying cyber bullying is depicted by the following the figure 1 below.

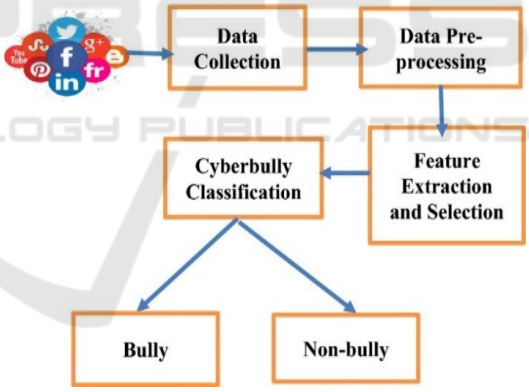

Figure 1: Social media types.

Dialectically, via social media, members of societies worldwide can communicate, exchange information, build personal relationships, and influence one another (Mozafari et al.; Mosca et al.). However, convenience brings risks, particularly among youth under 30, who face threats like online blackmail, cyberbullying, and emotional distress (Sabat et al.; Gravano et al.). Social networking platforms facilitate communication between individuals and organizations while simultaneously generating vast quantities of user-generated content (Alkomah & Ma). Among the most commonly applied methods to

Jayasurya, K., Periyasamy, T., Prasanth, S., Sibhisaran, S. and Ananya, A.
Leveraging Deep Neural Networks for Real-Time Detection of Cyberbullying and Offensive Memes in Social Network.
DOI: 10.5220/0013908200004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 4*, pages 69-74
ISBN: 978-989-758-777-1

analyze this content is sentiment analysis, widely used in previous research (Fortuna et al.; Aluru et al.).

1.1 Background

Social media operates as both a communication and an information-sharing platform due to its global reach (Patil et al.). However, the same digital connectedness allows harmful content like memes containing hate speech or provocative humor, and cyberbullying messages to spread rapidly (Toraman et al.; Antypas & Camacho-Collados). Cyberbullying involves deliberate acts of harassment or threats using digital tools, while offensive memes often relay derogatory and hateful messages with even greater virality (Sabat et al.; Zhou et al.). Such content affects users' mental health, undermines platform standards, and threatens the credibility of social media businesses (Velankar et al.; Akuma et al.). Given the volume of textual content online, it is infeasible for human moderators to review every post, even with filtering rules (Mullah & Zainon; Rabiul Awal et al.). This calls for real-time, automated detection systems to mitigate the spread of harmful material (Gravano et al.; Malik et al.).

2 LITERATURE REVIEW

This research applies a Deep Convolutional Neural Network (DCNN) to detect hate speech and offers a multilingual study using 16 datasets across 9 languages (Roy et al.; Aluru et al.). While hate speech has long been studied in humanities, its exploration in computer science remains nascent (Alkomah & Ma; Fortuna et al.). One key proposal is DeepHate, a deep learning model leveraging high-dimensional text representations for robust detection (Cao et al.). Another approach, BiCHAT, combines BiLSTM, CNN, and hierarchical attention for nuanced detection (Khan et al. 2022a), while the AngryBERT model jointly learns target and emotion contexts (Rabiul Awal et al.). A BERT-based transfer learning technique has also demonstrated success in detecting hate speech from social media platforms (Mozafari et al.).

The recently proposed HCovBi-Caps model integrates convolutional layers, BiGRU, and capsule networks to improve hate speech detection, particularly considering its abstract nature (Khan et al. 2022b). Further, methods like TF-IDF and Bag of Words have been evaluated in combination with machine learning models for hate classification on live tweets (Akuma et al.). The cross-domain transferability of hate speech detection models is examined by Toraman et al., while Antypas & Camacho-Collados evaluate robustness across datasets. Meanwhile, context-based methods (Mosca et al.) and sentiment-sharing models (Zhou et al.) have enhanced semantic understanding in hate speech systems. Moreover, offensive meme detection, focusing on pixel-level content, pushes hate moderation into multimodal dimensions (Sabat et al.).

3 EXISTING METHODOLOGIES

And even if individuals are communicating asynchronously over the Internet, social networking as a stylish and, most importantly, unobtrusive way for people to show their respect to each other's individual opinions and beliefs is already a part of everyday life. Today, the grown man who expresses hatred through politics, misogyny, and racial prejudice has harassed and mistreated other people. Social networks are also gaining popularity among users for online tyrant, including blackmail. Myspace is a common social networking site through which people can comfortably join one or many societies that they find interesting Due to the sheer amount of information that exists in SNSs, researchers have been able to conduct extensive research in a number of fields. Emotion analysis is one of the areas of research that receive a lot of attention and primarily use data from social media. Recommending filters arrange the offered content list in relation to the similarity between the desired item content and the end user's interests; whereas, collaborative filters offer items based upon comparison with other like-minded individuals. Since content-based filtering processes are more or less focused on text-based documents, Filtering, which categorizes the mass incoming texts into relevant and irrelevant, must truly be regarded as a single tag classification. Multi-label text categorization is an advanced form of filter that will categorize the communications partly by topics and categories. The classifier generates automatically according with the learning need from the set of previously classified sample in content-based filtering paradigm from the machine learning (ML) framework, identified that Bag of Words (BoW) method perform better when compare with complex text representation method that may have better statistical features but lower semantics.

4 PROPOSED METHODOLOGIES

Another common and well-known tool of information exchange share and communication involving personal data is considered to be an online social network (OSN). But one of the significant challenges for OSNs is to provide the users with control over their privacy and stopping the broadcasting of unsuitable content. This work provides a solution to the problem by allowing the users of OSN to decide on the messages that appear on their walls. This is done through allowing the users to set up their own specific filtering rules of their choosing in a highly flexible rule-based system. Moreover, there is utilizing of the soft classifier founded on the option of the machine learning for the messages' automatic recognition and content filtering. Every short message is analyzed by four Deep Learning (DL) text classification algorithms, which categorize the message into one or more groups based on its content. A dataset of the categorized words is developed to check for any words that should not be in any particular category.If there is any inappropriate language in the correspondence, it will be filtered to remove it from the Blacklists. Moreover, since the wall message uses material technique, no immoral word will be written on the wall of the user.

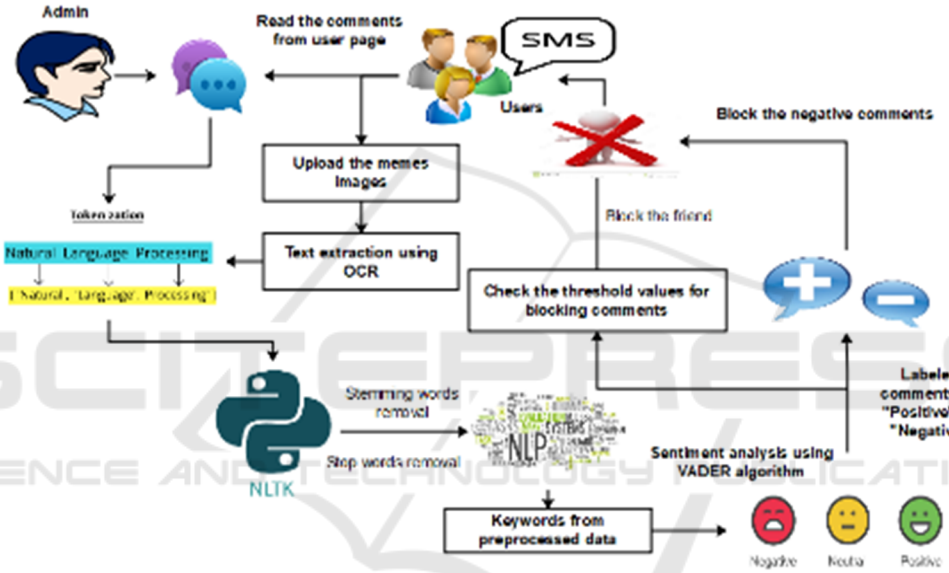

Figure 2: Framework.

Attributes such as text message content, users' buddy connections and other writer related characteristics are also considered by a system, utilizing blacklists for filtering contextless messages. The criteria of the proposed structure include the application of deep learning in helping the user define FRs used as an understanding of filter rules, enhancement of fit rules in the considered domain, and the expansion of the set of attributes to be analyzed initially by the classifier. context messages. The suggested framework incorporates deep learning techniques to assist users in defining Filtering Rules (FRs), improve fit rules within the domain under consideration, and broaden the collection of attributes initially examined during the classification process. Figure 2 shows the framework.

4.1 Framework

Social Networking Service or SNS is a particular type of online platform that makes it possible to keep links with people who have common interests, activities, backgrounds, or acquaintance for all people, regardless if it is casual or business related. As this point grows increasingly flexible, it has been very difficult to define social media. Social network is human relations that are in which people share, exchange and pass values, knowledge and experience. Such interactions can be enhanced by creating a graphical user interface where most of user communication and necessary graphical and visual triggers can be undertaken. Both user and admin interfaces may be designed using this module. We are providing users to interact with the system, respond to invitations, post photos with friends, and upload

meme images. They can then be forwarded to the admin page to determine and analyze.

4.2 Words Extraction

Social networking has recently become more and more an integral part of everyday online activities as social apps and websites are taking social networking forward as an important medium of communication. Today social networking sites allow people to do a lot of things that stimulate active participation such as leave user comments and share multimedia. But businesses know social media can help grow and increase visibility and use social media for marketing, brand promotion, connecting with customers as well as making new alliances. Online social networks that allow for free interaction and commentary are now open discussion, however, they need content management systems. By extracting text from photographs through optical character recognition (or text recognition), or OCR, content management, filtering, and analysis on these platforms will be improved.

4.2.1 Optical Character Recognition

- Extraction of Character boundaries from Image,
- Building a Bag of Visual Words (BOW) framework in remembering the Character images,
- Loading trained Model,
- Consolidating predictions of characters

The text may consist of bigrams, multigram, or unigrams. Feedback from social media users is gathered using this component. Links, text, and brief messages are just a few of the various formats that comments might take. After being reviewed, comments are sent to the server page.

4.2.2 Text Mining

Developing an effective deep learning classifier often emphasizes identifying and extracting key features that define and characterize the data. The typical process in text mining consists of these steps:

- Exclude commonly used words that do not add value to the analysis.
- Reduce words to their root forms to standardize variations.
- Remove symbols, punctuation, and other non-alphanumeric characters.
- Identify and isolate sentences that convey the most relevant information.
- Simplify or replace lengthy phrases with more appropriate or concise alternatives.

By taking these steps the text is pre-processed well for the deep learning models to effectively classify and analyse the text. By taking these steps the text is pre-processed well for the deep learning models to effectively classify and analyse the text.

4.3 Classification

In this module, we built an automatic mechanism which we called Filtered Wall (FW) to filter unsolicited communications coming from OSN user walls. The three-tiered architecture supports OSN services. The initial layer generally attempts to provide the basic OSN capabilities (relations and profiles). Additionally, some OSNs offer a further layer allowing us to deal with external Social Network Application (SNA). In the last, an additional layer may be required for supporting SNA's graphical user interfaces (GUIs). This thesis centres around the development of a robust backpropagation neural network (BPNN) that aims to extract and select a set of discriminative and characteristic features. Text classification interfaces between the constraints that are specified, and the constraints that will be applied in the process. In the approach, it is assumed that the easiest to remove (likely best) is "neutral" sentences, which do not contribute significantly to the classification task. Finally, the remaining 'non-neutral' sentences are assigned into the appropriate 'classes' of interest. This approach eliminates the burden of the work and allows us to process the data in a more organized fashion since, working from the VADER point of view, we suggest a hierarchical two-level method.

4.3.1 Developing the Deep Learning Model

The process works in developing a system to capture and remove any offensive content from user generated text. Here's the proposed workflow for creating a deep learning-based classifier:

- First, we have to start with a collection of a different kind of data or comments. This includes both examples of offensive (hateful) and non-offensive (neutral) language in this data.
- Remove any character, symbol, or irrelevant information that might interfere with analysis in cleaned collected text.
- Therefore, deep learning such as the VADER sentiment analysis algorithm or similar natural language processing (NLP) algorithms is used

to train a model to differentiate between hateful and non-hateful language.

- With the processed data feed into the model, you want to use and tweaking the model's parameters according to the labelled examples so the model can learn how to correctly classify hate speech.
- After training, run the model inside a system that can rapidly and correctly identify offensive language of new user inputs to keep things positive.
- Watchlists can also be used by the system to keep track of offensive terms or expressions; These lists auto flag content depending on the relationship and context with the user, their message. This screening can be tailored to the specific needs of the domain so that the filtering procedure is precise and appropriate in any situation.

By the use of deep learning methods to classify objectionable language, our solution ensures that objectionable content is still effectively controlled while increasing user experience.

4.4 Filtering Guidelines

Filtering should allow users to place restrictions on content creators to suit particular circumstances. Such guidelines can be based on conditions by characteristics represented in user profiles. For example, guidelines can be tailored according to the likes of managing them for young creators, people who espouse a certain religious or political view or those deemed to have started in a particular subject and lack experience. By applying filters to profile attribute, such as work-related information, it can be achieved. It can also use content-based parameters to identify, track and manage communication. Take the case of automatically banning people who keep posting negative reviews above a certain level — five times, let's say. Additionally, such actions should be notified to the users through those mobile devices.

4.5 Alert System

The BL overseeing system should be able to locate individuals in the BL and know when customer loyalty goals have been complete. The set of BL specific rules guiding this process is intended to increase system adaptability and threshold values are key to these rules. Having set the thresholds, the server defines acceptable levels of activity. This value can then be used by users to do actions like allowing or disallowing the posting of individuals that often comment in a critical way. In addition, the system should also send real time user notifications via smartphone notifications to maintain transparency and updated interfacing.

5 CONCLUSIONS

In this work, the hybrid model for harmful memes and cyberbullying in the social media environment is investigated, and we use the VADER sentiment analysis algorithm along with natural language processing to accomplish this. VADER helps the model recognize the subtle linguistic patterns often found in cyberbullying and is a proven effective processor of social media material leading to a reputable sentiment measure. Once OCR is added to image analysis, the combination of VADER can reliably detect offensive multimedia messages such as abusive memes containing grotesque text incorporated in images. The system's adaptability with respect to filtering options is also increased through management of BLs. This is the first stage of a bigger project. The early encouraging results with the classification technique have motivated us to work on other projects in order to improve classification quality. The DL soft classifier is used to filter out unwanted signals in this system. Using BL increases the filtering system versatility. We will design a system that incorporates more factors in determining whether or not to introduce a user into the BL. With a flexible language, the strong rule layer generates Filter Rules (FRs) that constrain the system to not display some data on their walls. The use in FRs is to utilize user profiles and relationships.

REFERENCES

Akuma, Stephen, Tyosar Lubem, and Isaac Terngu Adom. "Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets." International Journal of Information Technology 14.7 (2022): 3629-3635.

Alkomah, Fatimah, and Xiaogang Ma. "A literature review of textual hate speech detection methods and datasets." Information 13.6 (2022): 273.

Aluru, Sai Saketh, et al. "Deep learning models for multilingual hate speech detection." arXiv preprint arXiv:2004.06465 (2020).

Antypas, Dimosthenis, and Jose Camacho-Collados. "Robust hate speech detection in social media: A cross-dataset empirical evaluation." arXiv preprint arXiv:2307.01680 (2023).

Cao, Rui, Roy Ka-Wei Lee, and Tuan-Anh Hoang. "DeepHate: Hate speech detection via multi-faceted text representations." Proceedings of the 12th ACM Conference on Web Science. 2020

Fortuna, Paula, et al. "Directions for NLP Practices Applied to Online Hate Speech Detection." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022.

Gravano, Agustín, et al. "Assessing the Impact of Contextual Information in Hate Speech Detection." IEEE Access, vol. 11, pp. 30575-30590, 2023, doi: 10.1109/ACCESS. 2023.3258973. (2023).

Khan, Shakir, et al. "HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network." IEEE Access 10 (2022): 7881-7894.

Khan, Shakir, et al. "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection." Journal of King Saud University-Computer and Information Sciences 34.7 (2022): 4335-4344.

Malik, Jitendra Singh, Guansong Pang, and Anton van den Hengel. "Deep learning for hate speech detection: a comparative study." arXiv preprint arXiv:2202.09517 (2022).

Mosca, Edoardo, Maximilian Wich, and Georg Groh. "Understanding and interpreting the impact of user context in hate speech detection." Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media. 2021.

Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. "A BERT-based transfer learning approach for hate speech detection in online social media." Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8. Springer International Publishing, 2020.

Mullah, Nanlir Sallau, and Wan Mohd Nazmee Wan Zainon. "Advances in machine learning algorithms for hate speech detection in social media: a review." IEEE Access 9 (2021): 88364-88376.

Patil, Hrushikesh, Abhishek Velankar, and Raviraj Joshi. "L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models." Proceedings of the Third Workshop on Threat, Aggression and 2022.

Rabiul Awal, Md, et al. "AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection." arXiv e-prints (2021): arXiv-2103.

Roy, Pradeep Kumar, et al. "A framework for hate speech detection using deep convolutional neural network." IEEE Access 8 (2020): 204951-204962.

Sabat, Benet Oriol, Cristian Canton Ferrer, and Xavier Giro-i-Nieto. "Hate speech in pixels: Detection of offensive memes towards automatic moderation." arXiv preprint arXiv:1910.02334 (2019).

Toraman, Cagri, Furkan Şahinuç, and Eyup Halit Yilmaz. "Large-scale hate speech detection with cross-domain transfer." arXiv preprint arXiv:2203.01111 (2022).

Velankar, Abhishek, Hrushikesh Patil, and Raviraj Joshi. "A review of challenges in machine learning based automated hate speech detection." arXiv preprint arXiv:2209.05294 (2022).

Zhou, Xianbing, et al. "Hate speech detection based on sentiment knowledge sharing." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021.