INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078

(Volume 7, Issue 6 - V7I6-1422)
Available online at: https://www.ijariit.com

Detecting and preventing cyberbullying comments on social media using Deep Learning

M. S. Jothi
msjothibe@gmail.com
Jeppiaar Engineering College, Chennai, Tamil Nadu

Dr. J. Jospin Jeya
jeyasmith78@gmail.com
Jeppiaar Engineering College, Chennai, Tamil Nadu

ABSTRACT

Increasing social media sites attract the attention of all people. We can utilize its benefits effectively for good things. But few people don't use these sites in the right way. Hating comments, spreading rumours, blackmailing using social sites have been increased. Much research is involved in solving these problems by employing machine learning and deep learning algorithms. The most existing solution is detecting cyberbullying, but facing difficulties in analysing the short texts. Blocking the user who cyberbullying is also not possible. This paper proposes a detailed review of the machine and deep learning approach for detecting, preventing cyberbullying, and blocking malicious users using Natural Language Processing technique (NLP) and deep learning algorithm named as Back propagation neural network algorithm.

Keywords: *Cyberbullying, Detection, Prevention, Blocking Malicious Users, Natural Language Process, Back Propagation.*

1. INTRODUCTION

Nowadays, we all are traveling in our life with the internet. Especially in this pandemic situation, from school children to college students are efficiently utilizing this for their education purpose. Social media is a big part of social life and the creation of many young people and children. The benefits of social media include communication, learning, and creativity. Risks are inappropriate content, cyberbullying, and data breaches. The strong growth of social media users, especially survivors use, is associated with depression and various other side effects. In social media, users share their personal information. Social Networking Websites contain a large amount of text and/or non-text and so on information related to aggressive behavior. This leads to the growth of incidents of cybercrime, for example, cyberbullying has become a global epidemic. Examples of cyberbullying may include rumors posted on social media; embarrassing videos or pictures; and swearing, intimidation and harassing messages posted on social media. Cyberbullying bullying is frightening and devastating, which can lead to suicide attempts and causes lifelong brain damage to the victims. With the growing negative impact of cyberbullying on society, it is needed to detect and prevent these problems. Many researchers have borrowed their contributions to the development of methods in advance to identify and prevent it.

Hence, the aim of this project is to analyses the comments on social networks and block malicious users. Using the NLP technique can extract the keywords from user-generated content and implement a Back Propagation neural network to classify the text whether it is positive or negative. If it is negative means, automatically blocks the comments as per user wish and also blocks the friends based on predefined threshold values.

2. RELATED WORK

Author: P. Fortuna And S. Nunes, used Keyword selection and Recursive search techniques. They embrace a systematic approach that not only analyzes theoretical aspects but also the resources, such as data sets and other projects. This work also discusses the complexity of the concept of hate speech, which is defined in many forums and contexts and provides a concise overview. This place has an undoubted potential for social impact, especially in online communities and digital forums. The development and editing of shared resources, such as guidelines, multilingual data sets, and algorithms, is an important step in improving the automatic detection of hate speech. The advantage of their work is " annotating new messages and it was better than I. Alfina's work.

Author M. O. Ibrahim used Naive Bayes, Support Vector Machine techniques to detect abusive languages. But in their work, standard English words are alone detected.

Author R. CAO has proposed a deep reading framework known as DeepHate, which has used multilateral text presentations to detect hate speech automatically. Researchers have developed many traditional machine learning and in-depth learning methods to automatically detect hate speech on online social media. However, many of these methods consider only one type of text element, e.g., term frequency, or use of embedding. Such methods ignore some of the rich text information that can be used to improve the discovery of hate speech.

Author J. Salminen used Logistic Regression (LR), XGBoost (Extreme Gradient Boosted Decision Trees) to detect harmful comments in social media. The discovery of hate online is necessary to reduce the toxicity of social media. In this study, J.Salminen tested various machine learning models (Logistic Regression, Naïve Bayes, Support-Vector Machines, XGBoost, and Neural Network) to detect online hate and found the best performance with XGBoost as a separator with BERT features as a major influence. representation of hateful social media ideas. The fulfillment of the model on most social media platforms is good but varies slightly between platforms. Their findings support the goal of developing more hate online dividers on many social media platforms. It is a common perception that developing a hated class of the universe may benefit from the information retrieved from a variety of training sets and situations. The focus of the mono-platform is particularly troubling, as the lack of a common hate separator requires researchers and staff to "re-invent the wheel", meaning that each time they conduct online hate research (OHR) somewhere on the social media platform, the new classifier. it needs to be improved. This is successful not only in repeated intellectual efforts but also in "entry barriers" for researchers who do not have the skills to develop the model but who may be interested in interpreting OHR. Moreover, the shortage of class dividers worldwide means that results in all subjects and forums are not easily comparable. Overall, the diversity of models and the presentation of traits make it difficult to detect hate in all different forums and situations. The demerit of this approach is that it needs public available datasets.

3. PROPOSED SOLUTION

Online Social Networks (OSNs) today are one of the most popular collaborative methods of communicating, sharing, and distributing large amounts of personal health information. One important issue today on Social Networking Websites (OSNs) is giving users the ability to control messages sent to their private area to prevent the display of unwanted content. To date, OSNs provide limited support for this requirement. To fill the gap, in this paper, we propose a system that allows OSN users to have direct control over messages embedded in their walls. This is achieved through a flexible rule-based system, which allows users to customize the filtering conditions that will be used on their walls, and the soft Machine-based section automatically records messages that support content-based filtering. In-depth reading (DL) is used as a text-sharing technique to automatically share each short text message in a set of categories based on its content. Major efforts to build a robust back distribution algorithm focus on extracting and selecting a set with features that reflect discriminatory features and characteristics. Here, a classified dictionary is built and used to verify words that contain defective words. If the message contains any profanity, it will be sent to Blacklists to filter those words in the message. Finally, a message without defamation will be sent to the user's wall as a result of a content-based filtering process. The system automatically filters unwanted messages using blocklists based on both message content and message creator relationships. The main differences include the different semantics of filter rules to better fit the imaginary domain, to help users Define Filter Rules (FRs), to expand the set of features considered in the classification process.

4. SYSTEM ARCHITECTURE

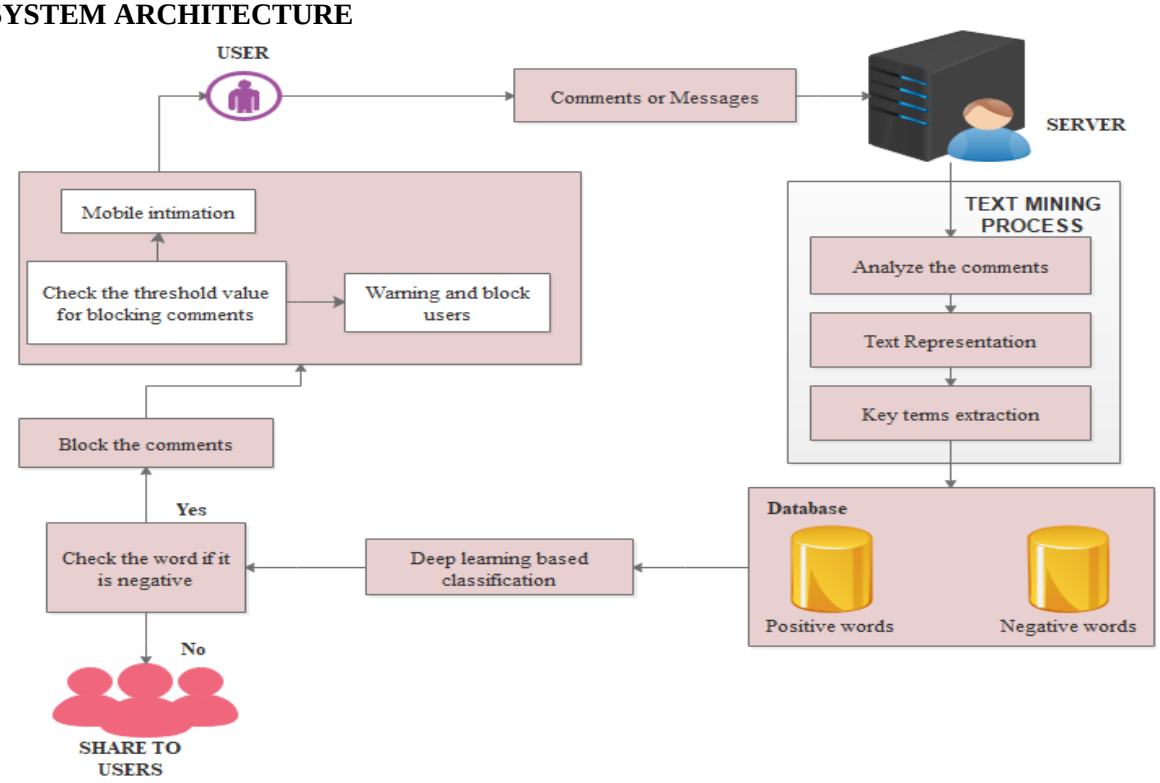

Figure a) System Architecture

Fig a represents the block diagram of the proposed work. The proposed work includes the user and server. The server can train the datasets with keywords related to positive and negative words. Users can comment on the page and apply the text mining steps to extract the keywords. Then classify the keywords using deep learning classifiers. If it is negative means, block the comments and also block the users based on threshold values with real-time mobile intimation. This project aims to analyze ideas in social media using the Indigenous Language (NLP) processing method and Deep algorithm called Backpropagation neural network algorithm. Using the NLP method, it can extract keywords from user-generated content and use the Back Propagation neural network to sort text whether it is good or bad. If it is a bad idea, automatically block comments as per user and block friends based on pre-defined values. Test results show that the proposed framework has been implemented in real-time for a social networking site with an improved notification system.

i) Natural Language Process :

Natural language processing (NLP) refers to the computer science department — and more specifically, the field of artificial intelligence or AI — which is responsible for giving computers the ability to understand the text and spoken words in the same way that humans can.

NLP combines computer languages — the official modeling of human languages — with mathematics, machine learning, and in-depth learning models. Combined, these technologies allow computers to process human language in the form of text or voice data and 'understand' the complete, complete meaning and purpose of the speaker or author and his or her mood. Human language is full of ambiguities that make it very difficult to write software that accurately determines the intended meaning of text or voice data. Synonyms, sarcasm, idioms, metaphors, grammar and usage, variations in sentence structure — these are simply grammatical errors that take years to learn, but that programmers must teach the use of natural language to recognize and understand well from the beginning if those apps are going to be helpful.

ii) Deep Learning :

Deep Neural Network (DNN) is a virtual neural network (ANN) with
multiple layers between input and output. DNN detects statistical deception in order to convert the input to output, whether in line or in offline relationships. The network walks the layers counting the chances of each output. For example, a DNN trained to identify dog breeds will go through a given image and calculate the probability that the dog in the image is a particular breed. The user can review the results and choose which opportunities the network should display (over a certain limit, etc.), and return the proposed label. Each mathematical illusion is considered a layer, and complex DNN has multiple layers, hence the term "deep" networks. DNNs can model non-linear relationships. DNN architectures produce design models where an object is presented as a horizontal structure of the original objects. Additional layers enable the formation of features from lower layers, which may be a more complex data model with fewer units than a shallow network that works in parallel. Deep structures incorporate a wide variety of a few basic techniques. Each building has achieved success in specific domains. It is not always possible to compare the performance of multiple properties unless tested on the same data sets. DNNs are usually frontal networks where data flows from the input layer to the output layer without backtracking. Initially, DNN creates a map of the visible neurons and assigns random numerical values, or "weights", to the connections between them. Weights and inputs are repeated and return the output between 0 and 1. If the network has not detected a particular pattern, the algorithm can adjust the weights.

5. IMPLEMENTATION

A) Framework Construction

A social networking service (also a social networking site, SNS or social networking
site) is an online platform that people use to build social networks or social networks that share similar personal interests or activities, domains, or real communication. A diverse and flexible range of standalone and built-in social networking services presents an explanatory challenge. A social network refers to the interaction between people where they create, share, and/or exchange information and ideas in virtual communities and networks. Design a GUI which is a type of user interface that allows users to interact with users using visual icons and visual cues. In this module, we can create a visual interface for administrators and users. The user can log in to the app and view a friend request. Users can share photos with their friends.

B) Read Comments

Social media is becoming an integral part of online life as social networking websites and apps proliferate. Most traditional online media sources include social media platforms, such as user feedback forums. In business, a social media platform is used to sell products, promote brands, and connect with current customers and promote new business. In this module, we can comment on the online social network. Mark in the form of text. Text can be uni-gram, bi-gram and multi-gram. This module is used to access input from social network users. Comments can be of various kinds such as links or texts or short texts. Comments are read and posted on the server page.

C) Classification

In this module, we design an automated system, called Filtered Wall (FW), which can
filter unwanted messages from OSN user walls. The structures that support OSN services are three-tiered structures. The first layer usually aims to provide basic OSN functionality (i.e., profile and relationship management). Additionally, some OSNs provide an additional layer that allows support for external network applications (SNA). Finally, supported SNAs may require an additional layer of communication required by the user (GUIs). Major efforts to build a strong back propagation neural network (BPNN) focus on the removal and selection of a set of discriminatory traits and traits. To clarify and enforce these issues, we use text classification. From BPNN's perspective, we approach this task by defining a two-tier sequence strategy by assuming that it is better to identify and remove “neutral” sentences and then classify “neutral” sentences by interest class in it instead of doing everything in one step.

6. CONCLUSION

In our project, we have designed a system to filter hateful messages from OSN
walls. The system exploits a DL soft classifier to enforce a customizable content-dependent filtered rules system. Great efforts to build a strong short text separator focus on extracting and selecting a set of discriminatory symbols and features. In addition, system flexibility regarding filter options is enhanced with BL management. This project is the first step in a comprehensive project. The encouraging early results we have achieved in the separation process urge us to continue another work that will aim to improve the quality of the separation. This system uses the DL soft classifier to remove unwanted messages. BL is used to enhance the flexibility of the system for filtering. We will be designing the system which will more sophisticated approach to decide when a user should be inserted into the BL. In addition to partitioning resources, the system provides a powerful legal framework that uses flexible language to define Filter Rules (FRs), through which users can specify which content should not be displayed on their walls. FRs can support different filtering processes that can be integrated and customized depending on the needs of the user. More precisely, FRs exploit user profiles, user relationships as well as the output of the DL categorization process to state the filtering criteria to be enforced. In addition, the system provides support for user-defined BlackLists (BLs), that is, lists of users that are temporarily prevented to post any kind of messages on a user wall.

7. FUTURE ENHANCEMENT

As a future project, we aim to use similar strategies to understand the rules of BL and FRs. In the future, we can extend the implementation framework of this program in a variety of languages with improved accuracy. It also includes a slow-track method of analyzing non-labeled data. Then add context comments that include emoticons to analyze comments on social media.

8. REFERENCES

[1] P. Fortuna and S. Nunes, ‘‘A survey on automatic detection of hate speech in text,’’ ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, Sep. 2018.
[2] Z. Al-Makhadmeh and A. Tolba, ‘‘Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach,’’ Computing, vol. 102, no. 2, pp. 501–522, Feb. 2020.
[3] R. Cao, R. K.-W. Lee, and T.-A. Hoang, ‘‘DeepHate: Hate speech detection via multi-faceted text representations,’’ in Proc. 12th ACM Conf. Web Sci., Southampton, U.K., Jul. 2020, pp. 11–20.
[4] Z. Waseem and D. Hovy, ‘‘Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,’’ in Proc. NAACL Student Res. Workshop, San Diego, CA, USA, Jun. 2016, pp. 88–93.
[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ‘‘Automated hate speech detection and the problem of offensive language,’’ in Proc. ICWSM, Montreal, QC, Canada, May 2017, pp. 15–18.
[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, ‘‘Deep learning for hate speech detection in tweets,’’ in Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion), Perth, WA, Australia, Apr. 2017, pp. 759–760.
[7] M. O. Ibrohim and I. Budi, ‘‘Multi-label hate speech and abusive language detection in Indonesian Twitter,’’ in Proc. 3rd Workshop Abusive Lang. Online, Florence, Italy, Aug. 2019, pp. 46–57.
[8] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, ‘‘Hate speech detection in the Indonesian language: A dataset and preliminary study,’’ in Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS), Jakarta, Indonesia, Oct. 2017, pp. 233–238
[9] M. O. Ibrohim and I. Budi, ‘‘A dataset and preliminaries study for abusive language detection in Indonesian social media,’’ Procedia Comput. Sci., vol. 135, pp. 222–229, Jan. 2018.
[10] J. Salminen, M. Hopf, S. A. Chowdhury, S.-G. Jung, H. Almerekhi, and B. J. Jansen, ‘‘Developing an online hate classifier for multiple social media platforms,’’ Hum.-centric Comput. Inf. Sci., vol. 10, no. 1, pp. 1–34, Dec. 2020
[11] A. Jha and R. Mamidi, ‘‘When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data,’’ in Proc. 2nd Workshop NLP Comput. Social Sci., Vancouver, BC, Canada, Aug. 2017, pp. 7–16.
[12] S. Yuan, X. Wu, and Y. Xiang, ‘‘A two-phase deep learning model for identifying discrimination from tweets,’’ in Proc. EDBT, Bordeaux, France, Mar. 2016, pp. 696–697.
[13] M. Mozafari, R. Farahbakhsh, and N. Crespi, ‘‘Hate speech detection and racial bias mitigation in social media based on BERT model,’’ PLoS ONE, vol. 15, no. 8, pp. 1–26, Aug. 2020.
[14] P. Burnap and M. L. Williams, ‘‘Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making,’’ Policy Internet, vol. 7, no. 2, pp. 223–242, Jun. 2015.
[15] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, ‘‘Detection of abusive language: The problem of biased datasets,’’ in Proc. HLT-NAACL, Minneapolis, MN, USA, Jun. 2019, pp. 602–608.

