Detection of Cyber bullying on Social Media Using MachineLearning

Vedangini Awate[1], Vaibhavi Bagad[2], Sharwari Jadhav[3], Bharti Jadhao[4]
Computer Engineering, Nutan Maharashtra Institute of Engineering and Technology/ Savitribai Phule PuneUniversity, India

****Corresponding Author***
Email Id:-*vedangini.awate2001@gmail.com*

ABSTRACT
With the widespread use of the Internet in modern times, a massive amount of data has been generated. However, along with its advantages, the cyberworld also has its own set of disadvantages. One such disadvantage is cyberbullying, which is an online crime. Cyberbullying occurs when bullying takes place online using technology. This research paper presents review of30 different cyberbullying researchers and the methods they use to detect bullying. Cybercrime refers to all criminal activities that use the Internet as a means of access and are carried out through computers, mobile phones, and other electronic devices. The previous research on detecting cyberbullying has been limited due to factors such as unavailability of the dataset, hidden identities of the predators, and privacy of the victims. To overcome these limitations, an efficient text mining method is proposed in this paper, which uses machine learning algorithms to actively detect bullying texts. The dataset used for evaluation is collected from myspace.com and Preverted-Justice.com. Unlike previous studies that only considered textual features, this study extracted three types of features, namely textual, behavioral, and demographic features from the dataset. The text features include intimidating words that could lead to real cyberbullying results. The behavioral trait observed is that if a person is bullied once, they might bully someone else later. The demographic features extracted from the dataset include age, gender, and location.The system's performance was evaluated using two classifiers, namely the Support Vector Machine (SVM) and the Bernoulli NB. The SVM classifier outperformed the Bernoulli NB with an overall accuracy of 87.14

Keywords:-Cyber Bullying, Support Vector Machine (SVM),sarcastic tweets, social media

INTRODUCTION

In recent times, cyberbullying has become a concerning trend that involves analyzing people's attitudes and opinions on social media platforms such as Facebook, Twitter, andblogs.

The primary objective of cyberbullying is to identify polarity, i.e., positive, negative, or neutral. Sarcasm is a specific type of cyberbullying that can reverse the polarity of a given text. Cyberbullying refers to sending misinformationto an individual or a community that causes heated debates among users.

The growth of social media platforms like Facebook, In- stagram, and Twitter has increased cyberbullying, making it common among teens. This aggressive behavior has a psychological and critical impact on the victim and can also be imitated by other members of the group. According to a new global data survey, 4,444 cyberbullying cases are on the rise every day.

Many young people spend their time online sharing information on social

networks. Social networks allow people to communicate and share information with anyone at any time. There are over 3.444 billion social media users worldwide. According to the National Criminal Security Council (NCPC), cyberbullying is an online activity that involves hurting or intentionally embarrassing another person through a cell phone, video game application, or other means of sending or sending a text, photo, or video. Cyberbullying can occur anytime and anywhere, and it is possible to reach anyone with internet access.

Texts, photos, or videos of cyberbullying can be published in unknown ways, and tracing their source can be difficult or impossible.

Moreover, it is often not possible to delete these messages later. The most popular websites for bullying on the internet are Twitter, Instagram, Facebook, YouTube, Snapchat, Skype, and a number of social networking sites. However, certain social networking sites like Facebook offer bullying prevention guidance, including a special section on how to report cyberbullying and prevent it from blocking users. Similarly, on Instagram, users can monitor or block individuals who share photos and videos that make them uncomfortable. Additionally, users can suggest improvements for the app and report violations to the community.

OBJECTIVES

The aim of the system is to detect and prevent cyberbullying on social media platforms by identifying and analyzing aggressive behavior and short-hand text in comment sections.

The ultimate goal is to alert users and generate a report containing details of the bully to help the government take action. Additionally, the system will track the number of bullying incidents and block the person responsible for bullying to prevent further harm to victims.

LITERATURE SURVEY

Paper Name:Identification and Classification of Cyber- crimes using Text Mining Technique

Author:Shiza Andleeb,Dr. Rashid Ahmed, Zaheer Ahmed, Maira Kanwal

Cybercrimes are all crimes done using the internet as an access medium and a piece of electronic equipment, like a computer or a mobile phone. The key problems restricting previous research in cyberbullying detection include a lack of datasets, predators' covert identities, and victims' right to privacy. In light of these elements, a successful text mining strategy utilizing machine learning algorithms is suggested to proactively identify bullying content.

The system's per- formance has been assessed using the dataset gathered from Preverted-Justice.com and Myspace.com. Contrary to a previ- ous study on the same dataset, which only took into account textual features, the current study extracts three different types of features from the dataset: textual, behavioral, and demographic information.

Textual characteristics include some bullying terms that, if used in the text, may outcome in actual cyberbullying. If a user engages in bullying once, they are more likely to engage in bullying in the future. While age, gender, and location are among the demographic characteristics collected from the dataset. Different performance measures for each of the two classifiers used to evaluate the system are utilized, and the Support Vector Machine classifier is determined to perform better than the Bernoulli NB with an overall accuracy of 87.14.[1]

Paper Name: A Bag-of-Phonetic-Codes Model for Cyber- Bullying Detection in

Twitter.
Author:Ankita Shekhar ,M. Venkatesan
These days, social networking sites like Twitter, Facebook, MySpace, and Instagram are establishing themselves as effec- tive communication channels. These are now an integral part of daily life. People may communicate their ideas and actions with those in their social circle, which helps them feel more a part of their community. However, there are limitations to this freedom of expression. On social media, people occasionally display their hostility, which damages the feelings of the people who are being attacked. Cyberbullying can be based on sexual orientation, race, or physical handicap.

As a result, effective surveillance is required to combat such situations. As a microblogging platform, Twitter observes online abuse on every single day. Tweets, on the other hand, are unpolished messages with numerous misspelled and cen- sored terms. In this research, a Bag-of-Phonetic-Codes model is proposed as a novel technique to identify cyberbullying. Correcting misspelled words and spotting restricted words can both be done by using word pronunciation as a feature. Correctly recognizing duplicate words may result in a vocab- ulary that is less, which would reduce the feature space. This suggested study takes its cues from the well-known Bag-of- Words paradigm for extracting textual properties. The Soundex Algorithm has been used to generate phonetic codes. In addi- tion to the suggested model, experiments with supervised and unsupervised machine learning methods on various datasets were conducted to comprehend the approaches and difficulties in the identification of cyber bullying. [2]
Paper Name:A Fairness-Aware Fusion Framework for Mul- timodal Cyberbullying Detection
Author:Jamal Alasadi,Ramanathan Arunachalam ,Pradeep K. Atrey, Vivek K Singh.
Recent allegations of prejudice in multimedia algorithms, such as lower face detection accuracy for women and people of color, have highlighted the urgent need to develop strategies that are equally effective for various demographic groups. As a result, we assert that a significant research problem is assuring fairness in multimodal cyberbullying detectors (e.g., equal performance regardless of the victim's gender). We suggest a fairness-aware fusion architecture that makes sure accuracy and justice are still crucial factors to take into account when integrating data from variousmodalities. The inputs from several modalities are integrated in this Bayesian fusion framework while taking into account the interdependencies between features and the various confidence levels connected to each feature. In particular, this approach distributes weights to various modalities depending on both their fairness and accu- racy. Results from using the framework to solve a multimodal (visual + text) cyberbullying detection problem show its valuein assuring fairness and accuracy.[3]

Paper Name:Name:Text Imbalance Handling and Classifi- cation for Cross-platform Cybercrime Detection using Deep Learning
Author:-Munipalle Sai Nikhila, Aman Bhalla , Pradeep Singh
Cyberbullying has recently grown to be a major problem. Everybody is being bullied online as internet, from politicians to 10-year-old children, so this is not only a problem for women. It is imperative to create a model using artificial
Intelligence that can identify bullying in cross-platform posts. However, the nature of textual datasets that are helpful for model creation is very unbalanced. Synonym substitution and artificial data generation using generative adversarial neural networks are the two main

approaches we suggest in this paper to address textual data imbalancing. With the aid of a classifier built on a convolutional neural network, we give a methodical study of our methods. Our research demonstrates how using generative adversarial network approaches to eliminate data imbalance before categorization enhances the overall perfor- mance of the model.[4]

Paper Name:Cyber Bullying and the Expected Conse- quences on the Students' Academic Achievement Author:Norah Basheer Alotaibi Cyberbullying is a recurring issue in the Saudi educational system that has been made worse by the advancement of digital technology and its pervasiveness inpractically every sphere of daily life. With this technology, it is not unexpected that harassment has spread to kids' online communities, where it is rife. Stakeholders have been worried by the incidence and results of this phenomenon, yet unexpectedly, studies exploring the reasons and motivations of cyberspace bullying involvement are scarce.

The Theory of Planned Behavior, a well-known theory, was used to explore this problem. (TPB). The impacts of attitudes, normative beliefs, subjective norms, and perceived behavioral control/self-efficacy were explicitly studied in this study. The motivations behind cyberbullying and

anticipated societal effects. This study offers important insights into students' intentions to engage in cyberbullying as well as the connections between Theory ofPlanned Behavior (TPB) factors and the predictive utility model. The results of this study can be used as a foundation for developing preventative and intervention measures, which has several implications for theory, practice, and policy.[5]

EXISTING WORK

In recent years, the issue of cyberbullying on social media has received significant attention, leading to an increase in research on automatic detection of cyberbullying. One notable study proposes a method for detecting cyberbullying in English textual data, which has received considerable attention in the research community.

The study approached the problem by dividing it into smaller text classification challenges, covering subtle topics, and collecting 4,500 text reviews from split YouTube videos. The study implemented Naive Bayes, Sup- port Vector Machines, J48 binary and multiclass classifiers, using broad and narrow functionality. Additionally, the study applied deep learning architecture to the Kaggle dataset, and conducted experimental analysis to evaluate the efficiency and performance of deep learning algorithms such as LSTM and BiLSTM .

ANALYSIS MODEL: SDLC MODEL TO BE APPLIED

A. Requirement gathering and analysis: In the initial phase of the waterfall methodology, we engage in requirement gathering and analysis. This crucial step involves identifying and documenting all the necessary requirements for our project. These may include software and hardware needs, database requirements, and interface specifications.

B. System design:

During the system design phase, ourfocus is on creatinga user-friendly system that is easy to understand for the end-user. We use various techniques such as creating UML diagrams and data flow diagrams to visualize the system flow, module interactions, and execution sequences. These design documents are essential to ensure that the system meets all the necessary requirements and is aligned with the stakeholders' expectations.

C. Implementation:

In the implementation phase, we focus on developing and integrating the various modules required to achieve the de- sired outcome. We use the system design documents as areference and break down the system into smaller units to be developed independently. These units are then integrated in the subsequent phases.

To ensure the quality of each unit, we conduct unit testing where each module is developed and tested individually for its functionality. This approach helps to identify and fix any errors early on in the development processand ensures that each unit works as intended.

D. Testing:

During the testing phase, we execute various test cases to verify whether each project module delivers the expected results within the expected time frame. After being individually tested, each unit created during the implementation phase is included into the overall system. Once all the units are integrated, the entire system is tested as a whole to identify and fix any possible defects or failures. This comprehensive testing process ensures that the system is stable, reliable, and meets all the necessary requirements before it is deployed for use.

E. Deployment of System:

Once the functional and non-functional testing is complete and any issues have been resolved, the product is ready for deployment. At this stage, the system is either deployed in the customer's environment or released to the market.

This involves the installation of the system in the customer's infrastructure or making the system available for download or purchase by end-users in the market.

The deployment phase marks the final step in the waterfall methodology, and it is important to ensure that the system is thoroughly tested and validated before deployment to avoid any potential issues or errors that may arise in the production environment.

F. Maintenance:

After the system is deployed, it enters the maintenance phase. In this phase, any issues that arise in the customer's environment are addressed, and patches or updates are released to fix these issues or improve the system's performance. The maintenance phase is critical in ensuring the long-term stability and reliability of the system.

It involves implementing changes and improvements in the customer's environment to keep the system functioning as intended. In the waterfall methodology, the various stages of the project are inter- connected in a cascading manner, and each phase must be completed and signed off before the next one can begin. The stages do not overlap, and progress flows smoothly down the stages like a waterfall, which gives the methodology its name.

PROPOSED SYSTEM

Twitter datasets are relatively easier to extract compared to other social media platforms such as Facebook, Instagram, and YouTube. Cyberbullying, according to statistical data, mostly occurs on Facebook, but since only public profiles can be easily accessed, data extraction is more manageable on Twitter where data is publicly available. The primary objectiveis to extract public data from social media platforms using available APIs. Subsequently, the extracted data is cleaned and preprocessed to enhance accuracy as the data contains multilingual, unstructured content as well as emojis.

To identify the best-suited classifier for cyberbullying detec- tion using content-

based features, several supervised machine learning algorithms were compared, and SVM was found tobe the most frequently used classifier among other supervised learning methods. SVM has been shown to be efficient in classifying heavily biased text, making it an excellent choice for cyberbullying detection.

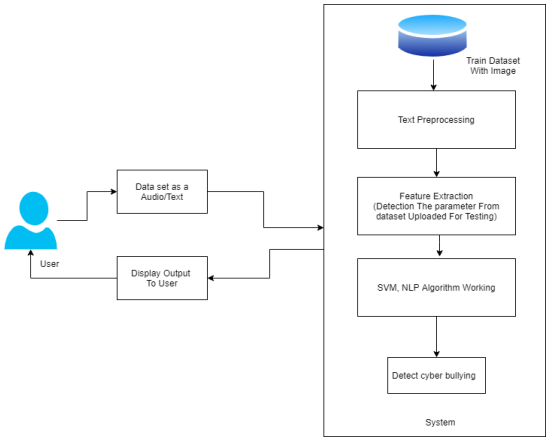

Fig.1:-*System Architecture*

FUTURE WORK

The system has been developed specifically to detect cyber- bullying in tweets. However, future work can be carried out to update and personalize the system as a real-time template, allowing users to check their Twitter accounts for cyberbully ing anytime and anywhere on their devices.

Additionally, the system can be modified to detect cyberbullying in photos andother social media platforms such as Facebook and Instagram. To enhance user experience, a module can be added to notify users of bullies identified by the model via email or SMS. A website can also be created where users can register and provide their Twitter IDs during registration.

CONCLUSION

This paper proposes a system for detecting cyberbullying in English and Hindi tweets on Twitter. The system relieson contextual cues, such as emotions and sarcasm, to detect cyberbullying. Instead of using a dataset, the system uses a collection of 9,104 cyberbullying tweets and the LR algorithm, which achieves good results and higher accuracycompared to other classifiers. However, traditional machine learning algorithms cannot handle the massive data generated by Web 4.0, and recent deep learning techniques, such as NLP, deep recurrent neural networks, CNN, and stacked autoencoders, may provide better results.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the pub lishers and researchers for providing their valuable resources, and also extend their appreciation to their instructors for their guidance. A special thanks goes to Professor Bharti Jadhao for her encouragement, guidance and support throughout this work. The authors are also grateful to Nutan Maharashtra Institute of Engineering and Technology, Pune for providing the opportunity to carry out this work.

REFERENCES

1. Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 8, pp. 1798–1828, 2013
2. A. M. Kaplan and M. Heinlein, “Usersof the world, unite! The challenges and opportunities of social media,” Business horizons, vol. 53, no. 1, pp. 59–68, 2010.
3. R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and
4. M. R.Lattanner, “Bullying in the digital age: A critical review and misanalysis of cyber bullying research among youth.”2014

5. B. K. Biggs, J. M. Nelson, and M. L. Sampilo, “Peer relations in the anxiety–depression link: Test of a mediation model,” Anxiety, Stress, Coping, vol. 23, no. 4, pp. 431–447, 2010.
6. K. Dinakar, B. Jones, C.Havasi, H. Lieberman, and R. Picard. ”Common sense reasoning for detect ion, prevent ion, and mit igat ion ofcyberbullying.” ACM Transact ions on Interactive Intelligent Systems (TiiS) 2, no. 3, 2012, p. 18.
7. V. Nahar, S. Unankard, X. Li, and C. Pang. ”Sentiment analysis for effective detect ion of cyber bullying.” In Asia-Pacific Web Conference, Springer,Berlin, Heidelberg, 2012, pp. 767-774.
8. V. Nahar, X. Li, C. Pang, and Y. Zhang. ”Cyberbullying detect ion based on text-st ream classification.” In The 11th Aust ralasian Data Mining Conference (AusDM 2013), 2013.
9. M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. ”Improving cyberbullying detect ion with user context

10. In European Conference on Information Retrieval, Springer, Berlin, Heidelberg, 2013, pp. 693- 696.