CYBER BULLYING DETECTION USING MACHINE LEARNING

HARIKA GUMMADAVELLY, KOMMAGONI PREETHI GOUD,

S KRISHNA KANTH REDDY

Computer Science and Engineering at Sreyas Institute of Engineering & Technology For

Jawaharlal Nehru Technological University, Hyderabad

N SANTOSH RAMCHANDER

Assistant professor

Computer Science and Engineering at Sreyas Institute of Engineering & Technology For

Jawaharlal Nehru Technological University, Hyderabad

ABSTRACT:

Cyberbullying is the act of sending threatening messages of somebody in order to insult them. It is difficult to prevent cyber victimization as a result of the activity. A cyber bully is someone who uses technology to bully another person. Cyberbullying and its impacts have occurred all over the world, and the number of cases is increasing. Cyberbullying detection is critical because online information is so large that it cannot be tracked by humans. This platform is to promote the Nave Bayes classifier for word extraction and loaded pattern

clustering. The goal of this study is to use Naive Bayes to build a classification model with the highest accuracy in identifying cyber bully conversations.

PROBLEM STATEMENT:

According to University of British Columbia research, cyberbullying is a bigger problem than traditional bullying. As per 733 adolescent surveys, 25- 30 percent of them had been involved in cyberbullying, while only 12 percent had been involved in traditional bullying. 95 percent of them stated that they only use mocks on the internet as a joke, whereas the rest are originally meant to insult or hurt someone. As per the report, teenagers vastly underestimate the dangers of cyberbullying.

OBJECTIVE:

Ability to detect cyberbullying through the use of online platforms is becoming increasingly important. Because there is too much information that humans cannot track, automatic detection that can identify threatening situations and hazardous content is required. This enables large-scale social media monitoring.

INTRODUCTION:

1.1 Introduction

2 Because of advancements in internet and information technology, Online Social Network (OSN) services such as Facebook, Twitter, and MySpace has become extremely popular as a main method of interacting with each other. Messaging is widely used and extremely useful for a variety of purposes, including business, education, and socialisation. However, it also allows the

development of harmful activities. There is a lot of evidence indicating that messaging can introduce a very serious problem, namely cyberbullying.

3 Cyberbullying is defined as the use of offensive information such as harassment, insult, and hatred in messages sent or posted via OSN services with the intent of intentionally injuring people emotionally, mentally, or physically [1]. It can lead to low self-esteem, anxiety, depression, and a variety of other emotional issues, as well as suicide [2]. Its tragic consequences have been repeatedly reported, most notably among school-age children. Because the number of cyberbullying incidents has recently increased [3,] an intensive study of how to effectively detect and prevent it from occurring in real time is urgently required. Blocking the message is not an effective way to protect victims from the incidents. Texts in messages, on the other hand, should be monitored, processed, and analysed as quickly as possible to support real-time decisions.

4 As a matter of fact of the issues raised, a number of studies are being conducted to investigate various techniques for effectively detecting cyberbullying. Manual detection is thought to be the most accurate, but it is rarely used because it takes too much time and resources. As a result, the automatic cyberbullying detection system is emphasised. The majority of text mining technology and techniques are used [4]. Karthik [5] used YouTube comments to train multi classifiers such as Nave Bayes, JRip, J48, and SMO. Vinita [6] extracted features from datasets from Kongregate, Slashdot, and MySpace using LDA and the weighted term frequency-inverse document frequency (TF-IDF) function. Homa [7] used an Instagram dataset to train a Support Vector Machines classifier, and Romsaiyud [8] used ExpectationMaximization (EM) to cluster documents from data streams for threat cyberbullying detection.

5 Despite extensive research into cyberbullying detection systems, cyberbullying remains a growing concern, and existing approaches are still inadequate, particularly when dealing with large volumes of data. Different types of OSN services can represent various data forms or patterns. Consequently, reducing computation time becomes critical. As a result, detecting cyberbullying remains difficult.

2. LITERATURE SURVEY

Several studies on cyberbullying analysis and detection using text mining by classifying conversations or posting have been published in recent years. Yin, Xue, and Hong [4] employ supervised learning, N-gram labelling, and TF-IDF weighting. Dinakar, Reichart, and Lieberman [5] used supervised machine learning to collect YouTube comments, manually label them, and implement various binary and multiclass classifications. Kelly Reynolds [6] labelled images using Amazon Mechanical Turk and the decision tree (J48) and k-nearest neighbour (k = 1 and k = 3). Because Support Vector Machines (SVM) are well proven in classification, Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, and Bart Desmet [7] use them as classification algorithms. In their research, when the preprocessing step occurs, they use the LeTs Preprocess Toolkit to apply tokenization, PoS-tagging, and lemmatization to the data.

Based on those facts, this study will be conducted to classify cyberbullying in text

conversations using a text mining method developed from Kelly Reynolds' previous research (2012). The study was carried out by identifying the characteristics of cyberbullying in the conversation as well as classification using SVM and Naive Bayes methods in comparison to Kelly Reynolds decision tree (J48) and k-NN (k = 1 and k = 3). Support Vector Machines (SVM) were chosen as the classification algorithm because they have been shown to perform well in high-skew text classification tasks. Even if Nave Bayes requires little data for training, it can produce excellent results. In addition to the classification of two classes performed by Kelly Reynolds, four classes and eleven classes will be classified in this research with the aim of making recommendations based on the classification results.

METHODOLOGY

This project will be built with Python and web technology. Within that, we will first search for and download the dataset needed to train the model. After downloading, we will pre-process the data before trying to transfer it to Tf-Idf. The dataset is then trained and the model is generated separately using the naive bayes, SVM (Support vector machine), and DNN algorithms. Then, using the FLASK framework, we will create a web-based application. We will retrieve real-time tweets from Twitter and then apply the generated model to these retrieved tweets to determine whether the text or images are cyberbullying or not. We use Python as the backend, Mysql as the database, and HTML, CSS, and javascript as the frontend.

3. TECHNIQUE OF DETECTION

3.1 Based on Text

Existing system:

• Cyberbullying keywords, pronouns, n-grams, Bags-of-words (BoW), Term Frequency Inverse Document Frequency (TFIDF), document length, and spelling content-based features are grouped. Content-based features are prevalent in our sample, with as many as 41 papers employing them. Because cyberbullying messages are frequently abusive and insulting, it is not surprising that profanity was discovered to be the most commonly used content-based feature across the reviewed studies, with 22 papers using the presence of profanity in text as an indicator for cyberbullying. Profanity lexicons were created by studies such as Dinakar et al. (2011), Perez et al. (2012), Kontostathis et al. (2013), Nahar et al. (2013), and Bretschnei-der et al. (2014) using wordlists compiled by the researchers or sourced from external libraries such as noswearing.com3 and urban dictionary.com. By equating the presence of profanity with cyberbullying, the use of the profanity lexicon salone overlooks other important aspects of cyberbullying, such as repetition and the presence of a power differential. Similarly, Rafiq et al. (2015) cautioned against using profanity as the sole feature for cyberbullying detection, arguing that not all use of profanity and cyber-aggression constitutes bullying.

3.2 Based on Text

While the majority of the research in our sample has focused on textual bullying, images and videos can also be used as delivery systems for

online bullying, and their impact can be just as, if not more, damaging. Furthermore, as social media platforms improve their ability to detect and prevent textual bullying, bullies are likely to turn to other forms of media to circumvent anti-bullying measures. Recent advances in image processing and OCR (Optical Character Recognition) make it possible to detect cyberbullying within media forms such as images, animations, and videos. With social media trends such as internet memes and viral videos becoming increasingly popular in recent years, bullies can easily exploit them to perpetrate cyberbullying. As a result, we believe that developing systems capable of detecting bullying content within multimedia files will be an important area for future research considerations.

4. IMPLEMENTATION

During the implementation phase, code is generated from the deliverables of the design phase, and is the longest phase of the software development life cycle. For a developer, this is the most vital stage of the life cycle because it is where the code is created. The implementation phase may overlap with the design and testing phases. There are numerous tools (CASE tools) available to automate the production of code based on information gathered and produced during the design phase.

5. SYSTEM ARCHITECTURE

The design phase's goal is to start organizing a solution to the problem, such as a necessity document. This section describes how the opening moves from the matter domain to the answer domain. The design phase meets the system's requirements. The design of a system is most likely the most important factor in determining the quality of the software package. It has a significant impact on the later stages, particularly testing and maintenance.

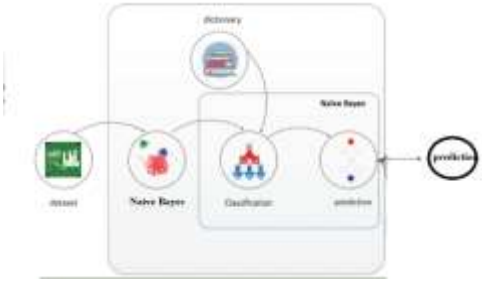
The style of the document is the result of this section. This document works similar to a blueprint of solution and is used later in implementation, testing, and maintenance. The design process is typically divided into two phases: System Design and Detailed Design.

System design, also known as top-ranking design, seeks to identify the modules that should be included in the system, the specifications of those modules, and how they interact with one another to provide the desired results.

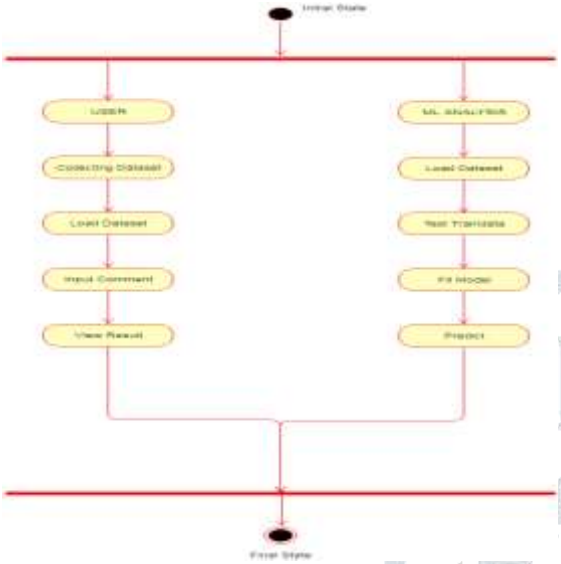
All of the main knowledge structures, file formats, output formats, as well as the major modules within the system and their specifications square measure set at the top of the system style. System design is the method or art of creating the design, components, modules, interfaces, and knowledge for a system in order to meet such requirements. It will be read by users because it applies systems theory to development.

The inner logic of each of the modules laid out in system design is determined in Detailed Design. Throughout this section, the fine print of a module square measure is sometimes laid out in a high-level style description language that is independent of the target language within which the software package will eventually be enforced.

The main goal of system design is to distinguish the modules, whereas the main goal of careful style is to plan the logic for each of the modules.

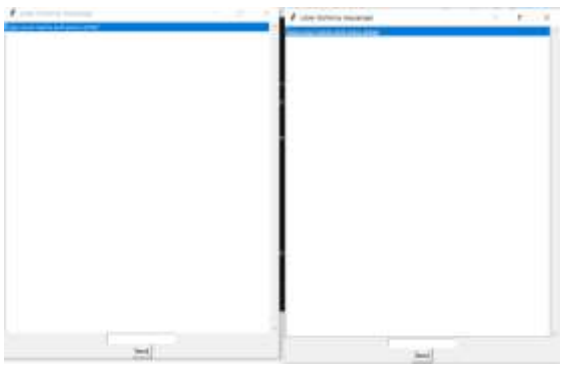


Architecture diagram

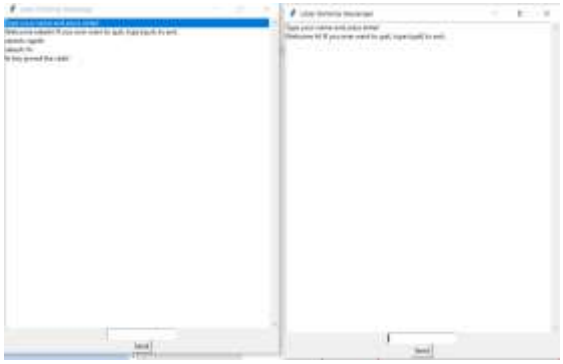


Here, we collect the data sets, process the data, and remove any impurities in the data sets. The data is then normalised if necessary so that it can be converted to a smaller volume of data. The data is then converted to a compatible format. And then it is stored in the databases. The required method is then used. We now have the final results.

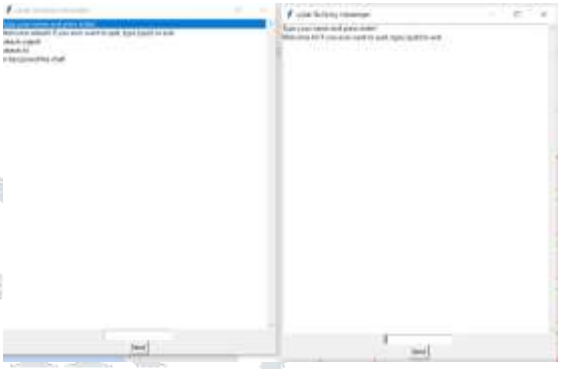
6. RESULTS



Chatting between 2 clients



When cyber bullying words are detected



7. CONCLUSION

The primary goal of this research is to improve the features of a Nave Bayes classifier for extracting words and generating models on text streaming. Furthermore, our proposed method ensures a local optimum. The method was tested on CyberCrime Data, a manually labelled dataset, for 170,019 posts and the Twitter web site for 467 million tweets. Because the data in this study are non-linear separable, the Poly kernel is the best Naive Bayes kernel for classifying cyberbullying, with an average accuracy of 97.11 percent. As a result, Naive Bayes with poly kernel is the best function for categorising the sample. Because of the highest accuracy level at n-gram 5 (92.75 percent), and the lowest accuracy set at n-gram 1, the use of n-gram may increase the accuracy level in cyberbullying classification (89.05 percent).

FUTURE ENHANCEMENTS:

- However, in the future, we will focus more on streaming K-mean clustering with Apache Spark to improve computation time and cost on various data types from large data sets.

References:

1. R.M. Kowalski and S.P. Limber, "Psychological, Physical, and Academic Correlates of Cyberbullying and Traditional bullying," J. Adolescent Health, 2013, vol. 53, no. 1, pp.513-520.

2. Cyberbullying Research Center, 'Summary of Our Cyberbullying Research (2004-2016)', 2016. [Online]. Available: http://cyberbullying.org/summary-of-our-cyberbullying-research. [Accessed: 10-Jul-2016].

3. M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context," Advances in Information Retrieval, Springer, 2013, pp.693-696.

4. D. Karthik, R. Roi, and L. Henry, "Modeling the detection of textual cyberbullying," International Conference on Weblog and Social Media - Social Mobile Web Workshop, 2011.

5. N. Vinita, L. Xue, and P. Chaoyi, "An Effective Approach for Cyberbullying Detection," Communications in Information Science and Management Engineering, 2013, vol. 3, no. 5, pp.238-247.

6. H. Homa, A. M. Sabrina, I. R. Rahat, H. Richard, L. Qin, and M. Shivakant, "Detection of Cyberbullying Incidents on the Instagram Social Network," 2015.

7. K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media (SWM'11), Barcelona, Spain, 2011.

8. I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. San Francisco, CA: Morgan Kauffman, 2005.

9. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kauffman, 1993.

10. W. W. Cohen, "Fast Effective Rule Induction," in Proc. Twelfth International Conference on Machine Learning (ICML'95), Tahoe City, CA, 1995, pp. 115–123.

ABOUT AUTHORS:

YERMOLU VENKATA DURGA DEVI. is currently pursuing MCA in SVKP & Dr K S Raju Arts & Science College, affiliated, to Adikavi Nannaya University, Rajamahendravaram. His research interests include Data Structures Web Technologies, Operating Systems and Artificial Intelligent.

PADALA SREENIVASA REDDY is working as an Associate in the Department of Computer Science in SVKP & Dr K S Raju Arts &Science College, Penugonda, A.P. He received MCA from Andhra University, 'C' level from DOEACC, New Delhi and M.Tech from Acharya Nagarjuna University, A.P. He attended and presented papers in conferences and seminars. He has done online certifications in several courses from NPTEL. His areas of interests include Computer Networks, Network Security and Cryptography, Formal Languages and Automata Theory and Object-Oriented programming languages.