International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

Cyberbullying Detection Using Machine Learning

Nideeksha B K[1], P Shreya[2], Sudharani Reddy P[3], Mohamadi Ghousiya Kousar[4]

[1,2,3]*B.E. Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, India*
[4]*Assistant Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology Bengaluru, India*

Abstract - *The advent of the digital age has enabled people to a new form of bullying which often results in social stigma. This new form of bullying is Cyberbullying which is a crime in which a perpetrator targets a person with online harassment and hate. Social networks provide a rich environment for bullies to find and harass vulnerable victims. Messages or comments concerning sensitive topics that are personal to an individual are more likely to be internalized by a victim, often ending in tragic outcomes. This phenomenon is creating a demand for automated, data-driven techniques for analyzing and detecting such behaviour on the internet. In this paper, a machine learning-based approach is proposed to detect cyberbullying activities from social network data. Naïve Bayes classifier is used to classify the type of message i.e., cyberbullying and non-cyberbullying message. Finally, a chatbot can be implemented to warn bullies about the consequences of their cyberbullying messages and take necessary actions. Our evaluation of performance results reveals that the accuracy of the proposed approach increases with more classification data.*

Key Words: Cyberbullying, machine learning, feature extraction, text classification, Naïve Bayes.

1. INTRODUCTION

Cyberbullying is a planned and repetitive act to harm or humiliate a person using information and communication technologies, including e-mails and social media. It is categorized into various forms, like cyber harassment (repetitively harassing and threatening someone), denigration/slandering (sharing false information about someone), flaming (brief insulting online interactions), etc.

Since the physical appearance of the bully is not required, it can go on nonstop. With the increasing adverse impact of cyberbullying on society, it's necessary to seek out ways to detect this phenomenon. Automatically recognizing emojis, bully words, and audio features from online social platforms, especially micro-blogging sites like Twitter, facebook and video-sharing platforms like YouTube is vital to research.

The process of detecting cyberbullying activities begins with input datasets from social networking sites. The input dataset consists of text comments and messages published on social media. Data pre-processing is performed on input data to enhance the quality of the research data. Subsequent analytical steps include removing extra characters, stop words, and hyperlinks. Feature extraction is done after completing pre-processing on the given input data. Features like Pronoun, Noun, and Adjective from the text are obtained by feature extraction and the frequency of words in the text is determined. The extracted features are then given as input to the Classification Algorithm.

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem The output of this classification algorithm shows if the given message contains bullying words or not. This is a supervised approach because it uses a tagged or labeled training data set.

2. LITERATURE SURVEY

In 2020, Alwin T. Aind et al. [1] novel algorithm Q-Bully can automatically detect cyberbullying using Reinforcement Learning and Natural Language Processing techniques. Cleaning consisted of stopwords, emojis, misspellings, and repeated letters. The Q-Learning Reinforcement Algorithm is used in the Q-Bully algorithm. The model would run over all of the terms in the sentence and give the sentence a large negative reward if it predicted the wrong direction; otherwise, it would give it a modest positive reward. The rewards fed back to past states using the Q-Learning approach are stored in the Q-Table. The Q-Bully algorithm efficiently converged for small datasets at a reasonable rate without any context given to words. The model was able to achieve 93% accuracy for 10,000 comments and 89.5% for the whole dataset.

In 2019, Peidong Zang et al. [2] proposed to build a Long Short-Term Memory Neural NetworkDeterministic Finite Automaton (LND) model which considers not only the language content, but also the user's characteristics and historical speech on social network. Due to the lack of labeled content Douban's reviewer's data was utilized by analysing speech patterns with polarized emotions. Then the learned model was applied to analyze Chinese cyber bully behaviours on Weibo. As a result, the accuracy of detecting cyber bullying increased from 89% (sensitive lexicon filtering method) to 95% by considering user's behavior features and language emotional polarity scores.

In 2019, Ong Chee Hang and Halina Mohamed Dahlan [3] proposed exclusion cyberbullying lexicon by using an ontological approach. This approach consists of several

phases: understanding the concepts of exclusion cyberbullying, word list selection, keyword identification, classes, and subclasses identification, and lastly cyberbullying ontology and lexicon development. The selection of the exclusion cyberbullying keywords is based on some criteria, such as, the words that mean to (i) excluding someone, (ii) expression of dislike someone. The identified factors of exclusion cyberbullying were mapped with the related semantic frame in FrameNet. The selected exclusion cyberbullying keywords or lexical entries in FrameNet will be used as the subclasses. Next, the documentation is inserted into a subclass. OILEd and Protégé have been chosen as the ontologies-related tool to create, view, manage and edit the ontology, while OIL+DAML was chosen as the ontology language, and FrameNet as the English lexical database that provides definition, annotated sentences, and semantic frames.

In 2019, Jianwei Zhang et al. [4] proposed an optimal model for automatic detection of cyberbullying based on extracting different textual elements and examining their effects using multiple machine learning models. Textual features are extracted using n-gram, Word2Vec, Doc2Vec, emotion values of tweets, and Twitter-specific characteristics. The collected tweets are divided into training data and test data, and the models are constructed on the training data using each type of feature and each type of machine learning algorithm. The machine learning algorithms include linear models (Linear support vector machine, Logistic regression), tree-based models (Decision tree, Random Forest, Gradient boosting regression tree), and deep learning models. With the experiments based on the collected tweets, the quality of automatic cyberbullying detection is evaluated and the best model performs over 90% for the four criteria: accuracy, precision, recall, and F-measure.

In 2018, Hani Nurrahmi and Dade Nurjanah [5] proposed a cyberbullying detection method for the Indonesian language. The first step in cyberbullying tweet detection is pre-processing-Data cleaning, Tokenization, and POS Tagging. Then system will extract the features in the preprocessed text. Then, system will learn the pattern of the features using KNN and SVM (linear and RBF kernel). The system evaluates each class based on the learning model f-score. After that, the system calculates the behavior of the user steps is as follows: First, Normal Behavior (NB), calculated from the total number of Non-Cyberbullying tweets sent by user X in time (NB(X, t)). Second, Abnormal Behavior (AB), calculated from the total number of Cyberbullying tweets sent by user X in time (AB(X, t)). Third, Out-degree, calculated from the total of the tweet that sent from user X to other users in time t. Fourth, calculate the probability of NB (PNB) and AB (PAB) using PNB = NB(X, t) NB(X, t) + AB(X, t) and PAB = 1 − PNB.

In 2018 Hugo Rosa et al. [6] proposed to detect cyberbullying instances in unbalanced datasets and compare how Fuzzy Fingerprints perform. Three versions of the dataset are used which include a down-sampled balanced version, a full version with the unbalanced class ratio, and a balanced training dataset. The following are Fuzzy Fingerprints' parameters eligible for optimization, and their possible values: size k of the fingerprint: {20, 100, 500, 1000, 2500,5000}, minimum word length: {1, 2, 3}, threshold of the T2S2: {0.1, 0.25, 0.5}, removal of stop words: {YES; NO}, calculate the Inverse Class Frequency (ICF): {YES; NO}. FFP achieved an f-measure = 0.425.

In 2018, Daphney–Stavroula Zois et al. [7] suggested a novel algorithm AvOID for optimal online cyberbullying detection is used. Consider, for example, the case of two features f(m) = {y1, y2}, where y1 is the number of bad words, and y2 is the number of exclamation marks in a message. The prior probability p of a message being an instance of cyberbullying is first set to the posterior probability 0 of a message being an instance of cyberbullying, and the two terms are compared. If the first term is less than or equal to the second term, AvOID stops and classifies the message based on optimal strategy. When the first term is greater than the second, the first feature from the message is retrieved and assessed; as a consequence, a comparison outcome y1 is generated and utilized to update the posterior probability 0 to 1 using the update rule.

In 2018, Batoul Haidar et al. [8] proposed a cyberbullying detection method for the Arabic language. The Arabic dataset was extracted from Twitter and annotated manually. Firstly, all hyperlinks, non-Arabic characters, and emoticons were removed. Label "0" for non-bullying and "1" was used for bullying content. The dataset was tokenized into words eliminating all unneeded characters. After that word embeddings were created using one-hot encoding. The model was trained to employ a Feed-Forward Neural Network on this Arabic dataset. The FFNN model was developed with 4 hidden layers. The model was configured to separate the dataset into 80% training and 20% testing and shuffle the dataset at each epoch. This gave a performance metric of 91.17%. it had been found that validation accuracy and test accuracy were high when 7 hidden layers were used. Another parameter affecting performance is that the batch size, optimal being 16 for the used network.

In 2017, Noviantho et al. [9] constructed a classification model with optimal accuracy in classifying cyberbully conversation using the Naive Bayes method and Support Vector Machine (SVM) then applying n-gram 1 to five for the number of classes 2, 4, and 11 for every method. After Data Collection, preprocessing was done. For the needs of data balancing on the classification of two classes, 4 classes, and 11 classes based on severity level. The preprocessing text conversations are going to be transformed into a vector space model where text conversations are represented with a vector of extracted features. Features resulting from the extraction are words or combinations of words to make a list of words and therefore the calculation of the weight with TF-

IDF. The classification will use the SVM and Naïve Bayes method with linear, poly, RBF, and sigmoid kernels. the foremost optimal SVM kernel is that the Poly kernel (97.11%) and therefore the highest accuracy level is at n-gram 5 (92.75%).

In 2017, Elaheh Raisi and Bert Huang [10] introduced an algorithm that uses the topology of the communication network to learn a relational model. The algorithm looks for a consistent parameter setting for all users and key phrases in the data that characterizes each user's inclination to harass or be harassed, as well as a key phrase's tendency to be suggestive of harassment. The parameters are optimized by the learning algorithm to reduce their disagreement with the training data. The algorithm discovers new terminology. This feedback loop is repeated. The algorithm takes under consideration the entire communication network, propagating its bullying role estimates through the messaging structure and language utilized in each message.

In 2017, Walisa Romsaiyud et al. [11] proposed a novel method that can generate a predictive model from a large volume of data sets for supporting the analysis services on business. The first step, preparing the dataset. In the second phase, clustering, data sources have clustered the features of two types of communications: polite messages and abusive messages, in which the contents of the messages are recognized using the Kmeans clustering technique, based on a crime pattern and normalized documents. In the third phase, the abusive partition is utilized to transform each training data into a feature extractor for the Naive Bayes classification approach. The feature sets are fetched into the model in the final stage, predicting data, to yield projected labels.

In 2015, B.Sri Nandhini and J.I.Sheeba [12] proposed a framework for detecting cyberbully activities, which includes the following steps - Data Pre-processing, Feature Extraction, FuzGen learning algorithm, and Naive classifier technique. Learning algorithm unit consists of a genetic algorithm. Knowledge is represented by a collection of fuzzy rules. The major function is to change the way information is represented for classification while keeping the fundamental knowledge from the past. This information is stored in a chromosomal population that is processed by a genetic algorithm. All of the chromosomes in the population are vying to predict how cyberbully acts are classified. The results from the learning unit is assigned to the Classifier technique classifies the cyberbully activities using the fitness value of the chromosome.

3. METHODOLOGY

3.1 Dataset

The training dataset used consisted of 1066 samples contained in a CSV File. Each sample is a sentence followed by the corresponding target label. The target label is 'pos' for a NonCyberbullying Sentence and 'neg' for a Cyberbullying Sentence. The data is then fed to the model for training.

3.2 Building Classification Models

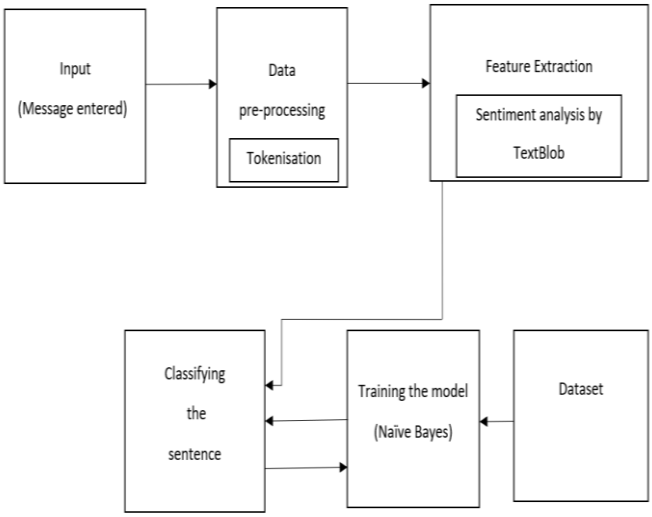
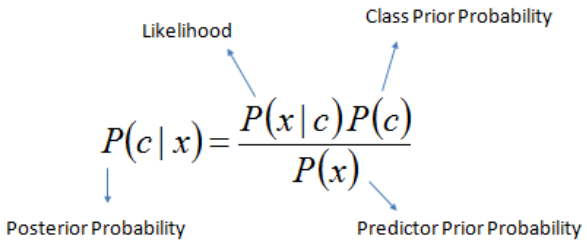


Fig -1: System Architecture

After the dataset is collected, a combined classifier is created. The values from Naïve Bayes Classifier model, Polarity of the phrase (Sentiment Analysis), and Vulgar confidence are then passed to a Combined classifier which combines results from other three classifiers and returns the appropriate Cyberbullying confidence level, which is then displayed by a bot. This customer classifier model can be further trained.

A. Naive Bayes Model

The Naive Bayes method is a supervised learning algorithm that solves classification problems and is based on the Bayes theorem. It is mostly used in text classification problems that necessitate a large training dataset. It's a probabilistic classifier, it makes predictions based on an object's probability. After the dataset is collected, a Naive Bayes classifier is created. The created model is then stored in a pickle file so that the file can be restored to update the model. We then use pickle.dump() to dump the data. It classifies text by reading from the pickle file and returns a float value between 0 and 1, where 0 if it is a positive message and 1 if negative.
Naive Bayes theorem can be used to calculate the posterior probability P(c|x) using P(c), P(x), and P(x|c).



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

The probability of the message 'd' being in class 'c' is computed as follows

$$P(c \mid d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k \mid c)$$

Where, $P(t_k \mid c)$ is the conditional probability of term t_k occurring in a message of class c.

B. Sentiment Analysis

TextBlob, a python library that aims to provide access to common text-processing operations, is used for Sentiment Analysis. When a user sends a message, the message is checked for the phrase's polarity. The parameter is the phrase which has to be checked for profanity. The polarity will be any float bounded by -1 and 1 where 1 indicates the sentence is most negative and -1 indicates its most positive.

C. Vulgar Confidence

We import the mysql.connector because python needs a MySQL driver to access the MySQL database. The database contains a table named badwords containing a list of bad words. The list of badwords is fetched from the database and stored in a list. The function checks if a word is in the swear list and returns the vulgar confidence which is -1 if there are no bad words and 1 if there are bad words.

3.3 Improvement of the Classification Models

When a user sends a message which is classified as a Cyberbullying message. The bot warns the bully and reports it to the reporting channel automatically. Furthermore, users with a specific role are also able to manually report the message by using the id. In the administration channel, they are able to thumbs up and thumbs down. If a user thumbs up, the offensive message will be deleted. Furthermore, the custom server classifier will be trained using this data. If the reaction is a thumbs up, the classifier is trained with a positive label else with a negative one.

4. RESULTS

The accuracy of this custom classifier will be poor in the server's infancy, but it will improve as more classification data is collected. The bot ignores a non-bullying message that is recognized correctly. When a cyberbullying message is detected, the bot warns, displays the confidence level, and deletes it. When a person sends a cyberbullying message that isn't classified correctly, users with a specific role can manually report it and give it a thumbs-up reaction. The message will be deleted, and hence the classifier will be trained with this example. When the user sends the same message again the message is predicted with a higher cyberbullying confidence level. When a user sends a non-cyberbullying message that is detected as cyberbullying, a thumbs down reaction can be given, and the classifier will be taught using this example so that when the user sends the same message again, it will be detected positive. Hence, the accuracy of the model increases with more usage.

5. CONCLUSION

Although social media platform has become an important entity for everyone, cyberbullying has many negative impacts on victim's life which include depression, anxiety, anger, fear, trust issues, low self-esteem. Therefore, detection of cyberbullying in the vast social media network has become immensely important. In summary, we have studied cyberbullying detection for messages and chats to identify cyberbullying text and actors on discord Platform. The work has successfully identified cyberbullying messages using the Naive Bayes algorithm.

6. FUTURE SCOPE

In future studies, we would like to add more dataset to improve the classifier accuracy. We would also like to implement the proposed approach to detect cyberbullying in several languages as social media is vast and isn't restricted to one language. There has been significant work done in text-based cyberbullying detection, yet working on audios/ videos cyberbullying detection remains a challenge.

REFERENCES

[1] A. T. Aind, A. Ramnaney and D. Sethia, "Q-Bully: A Reinforcement Learning based Cyberbullying Detection Framework," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-6, doi: 10.1109/INCET49848.2020.9154092.

[2] P. Zhang, Y. Gao and S. Chen, "Detect Chinese Cyber Bullying by Analyzing User Behaviors and Language Patterns," 2019 3rd International Symposium on Autonomous Systems (ISAS), 2019, pp. 370-375, doi: 10.1109/ISASS.2019.8757714.

[3] O. C. Hang and H. M. Dahlan, "Cyberbullying Lexicon for Social Media," 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS), 2019, pp. 1-6, doi: 10.1109/ICRIIS48246.2019.9073679.

[4] J. Zhang, T. Otomo, L. Li and S. Nakajima, "Cyberbullying Detection on Twitter using Multiple Textual Features," 2019 IEEE 10th International Conference on Awareness

Science and Technology (iCAST), 2019, pp. 1-6, doi: 10.1109/ICAwST.2019.8923186.

[5] H. Nurrahmi and D. Nurjanah, "Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility," 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 543-548, doi: 10.1109/ICOIACT.2018.8350758.

[6] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro and L. Coheur, "Using Fuzzy Fingerprints for Cyberbullying Detection in Social Networks," 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2018, pp. 1-7, doi: 10.1109/FUZZ-IEEE.2018.8491557.

[7] Zois, Daphney-Stavroula & Kapodistria, Angeliki & Yao, Mengfan & Chelmis, Charalampos. (2018). Optimal Online Cyberbullying Detection. 2017-2021. 10.1109/ICASSP.2018.8462092.

[8] Haidar, B., Chamoun, M., & Serhrouchni, A. (2018). Arabic Cyberbullying Detection: Using Deep Learning. 2018 7th International Conference on Computer and Communication Engineering (ICCCE), 284-289.

[9] Noviantho, Isa. S.., & Ashianti, L. (2017). Cyberbullying classification using text mining. 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), 241-246.

[10] E. Raisi and B. Huang, "Cyberbullying Detection with Weakly Supervised Machine Learning," 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2017, pp. 409-416.

[11] W. Romsaiyud, K. na Nakornphanom, P. Prasertsilp, P. Nurarak and P. Konglerd, "Automated cyberbullying detection using clustering appearance patterns," 2017 9th International Conference on Knowledge and Smart Technology (KST), 2017, pp. 242-247, doi: 10.1109/KST.2017.7886127.

[12] Nandhini, B. & Immanuvelrajkumar, Sheeba. (2015). Online Social Network Bullying Detection Using Intelligence Techniques. Procedia Computer Science. 45. 485-492. 10.1016/j.procs.2015.03.085.