**Detection of Cyberbullying on social media using Machine Learning**

By

Dileep Raja Chinni

**Abstract:**

The exponential rise in internet usage by people from all cultures and educational levels has made toxic online content become a significant problem in today's society. The automatic detection of toxic text content has significant challenges in differentiating between hate speech and offensive language. In this study, we suggest a method for classifying text into two categories: hate speech and non-hate speech. This project is focused on the detection of offensive data, and bully statements in shared data of social networks. In this project we will use the ‘Hate Speech and Offensive Language Dataset’ which has 31,935 tweets. The dataset was heavily skewed or imbalanced with 93% of tweets or 29,695 tweets containing non-hate labelled tweets and 7% or 2,240 tweets containing hate-labeled tweets data. Automated detection of these incidents requires the use of sophisticated and intelligent systems. Existing classical machine learning algorithms such as Random Forest, SVM, Naive Bayes with text cleaning for detecting cyberbullying have a lot of bottlenecks with less accuracy i.e., around 90%. We show how state-of-the-art deep learning NLP language models can overcome these bottlenecks with better accuracy and semantic understanding with hate and no-hate tweets.

**Introduction:**

Bullying or harassment that takes place online is known as cyberbullying or cyberharassment. Online bullying also refers to cyberbullying and cyberharassment. As the digital environment has grown and technology has improved, it has become more and more prevalent, especially among youth [1]. In many circumstances, cyberbullying is a continuation of existing traditional bullying. Most often, students who are bullied online have also been bullied in more conventional ways in the past (e.g., physically, or verbally). Few children experience bullying completely online; because these cybervictims are frequently physically larger pupils, bullies favor online confrontations over in-person interactions at school [2]. According to other studies, cyberbullying increased during the COVID-19 pandemic, when many children and adults were isolated at home and consequently spent more time online than they had before the pandemic.

In the Journal of Social Psychology, for instance, a study of adults found statistically significant increases in both pro-cyberbullying views and cyberbullying offending behavior [3]. Internet trolling is a popular form of bullying that occurs in an online setting (such as social media or online gaming) in an effort to cause a disturbance or to entertain the perpetrator. Another type of bullying or harassment is cyberstalking, which involves following a person through electronic contacts and may be a real threat to them. Cyberstalking is a type of internet harassment where a victim is stalked using electronic communications. Because there is typically a real danger to the victim's safety, this type of cyberbullying is regarded as being more hazardous than others. Cyberstalkers may send repeated communications with the intention of threatening or harassing people, and they may also overtly encourage others to do the same by acting like their victim and asking for contact information [4].

Cyberbullying can take place on social networking platforms such as Facebook, Myspace, and Twitter. "By 2008, 93% of young people between the ages of 12 and 17 were online. Aside from sleeping, youth spend more time with media than any other activity. Cyberbullying, which is defined as bullying that takes place using electronic communication technologies, such as e-mail, instant messaging, social media, online gaming, or through digital messages or photos sent to a mobile phone, has increased over the past ten years. Due to the abundance of technology available to kids today, cyberbullying has increased in prevalence. Instagram, Twitter, and Snapchat are the three most popular apps used by youngsters to engage in cyberbullying. Because parents and teachers are ignorant of when and where it is occurring, it has become more difficult to halt cyberbullying. According to 2006 research, 30% of adolescents and 45% of teenagers experienced cyberbullying at school. The reason pupils had access to their online devices, such as smartphones or computers, was related to this. Teens will say horrible things to one another online, but they don't aware that once they say it and put it online, it is there forever.

But in the current social sites do not focus on services like tracking the user behavior of anonymous behavior. In current system, social network sites need to focus the user microblogs and need to capture the user behavior whether his/she anonymous user or not. Few surveys’ providing concepts to tracking the attackers like using profile matching techniques and network-based techniques etc. But in real-time to apply those concepts in social network is less practical. Crawling the user information from the user micro blogs is also less practical. Anonymous Users can easily manipulate the public profile information. In social networks user may share their messages by using the chat applications. For every social networking sites has their own chat applications, for this Facebook is main example. And another way is sharing the multimedia data like images or videos. For this Facebook and Instagram best examples. For communication between users chat applications will be most useful for share their information, thoughts, views etc. But in the same way it may also cause the security loophole of user’s security which is cyber bullying. Such text-based content may security threat to the users because of the people can share cyber bullying words to the users with their fake accounts. [5], [6].

Machine learning (ML) is a topic of study focused on comprehending and developing "learning" methods, or methods that use data to enhance performance on a certain set of tasks. It is a component of artificial intelligence. Without being expressly taught to do so, machine learning algorithms create a model using sample data, also referred to as training data, to make predictions or judgments. Machine learning algorithms are utilized in a wide range of applications, including speech recognition, email filtering, computer vision, and medicine, when it is challenging or impractical to create traditional algorithms to carry out the required functions. Computational statistics, which focuses on making predictions with computers, is closely related to a subset of machine learning, but not all machine learning is statistical learning. The field of machine learning benefits from the tools, theory, and application domains that come from the study of mathematical optimization.

Data mining is a related area of study that focuses on unsupervised learning for exploratory data analysis. Some machine learning applications employ data and neural networks in a way that closely resembles how a biological brain function. Machine learning is also known as predictive analytics when it comes to solving business problems [7]. Although "deep" machine learning can use labeled datasets, commonly referred to as supervised learning, to guide its algorithm, it is not necessary. It can automatically identify the attributes that separate several types of data from one another and can ingest unstructured material in its raw form (such as text and photos). We can scale machine learning in more exciting ways since it doesn't require human intervention to handle data, unlike machine learning. Progress in fields like computer vision, natural language processing, and speech recognition is mostly attributed to deep learning and neural networks.

**Existing System:**

The availability of communications, multi-media, and e-commerce services has significantly risen because of the emergence of several social networking sites in the current trend. For instance, Twitter, a social media site with more than 700 million users, provides extensive microblogging services on a massive scale. Each day, 400 million microblogs are created on Twitter. On every social media network, including Twitter, Facebook, Sina, and others, there are more than 30% of fake, duplicate, or fraudulent accounts. The current social media networks, however, do not give services like anonymous user activity tracking a high priority. Whether a user is anonymous or not, social networking services today must focus on user microblogs and track user behavior. Few surveys’ providing concepts to tracking the attackers like using profile matching techniques and network-based techniques etc. But in real-time to apply those concepts in social network is less practical.

**Related work:**

The current work on detecting cyberbullying has focused on analyzing the text's features and detecting abusive behavior. In general, text classification, emotional analysis, and other technologies were used in combination with text content analysis. A few methods for extracting text features, such as N-gram models, bags-of-words models (Bow), TF-IDF, etc., were also successful [9]. Classification techniques, such as support vector machine (SVM), Logistic Regression, Naive Bayes, Random Forest, etc., have a significant impact in response to these features [15]. In order to identify cyberbullying, Chavan et al. [16] extracted several features (such as TFIDF, Bow, bullying words) and used SVM and Logistic Regression. Network factors have also drawn attention in addition to text characteristics [14]. The researchers also have evaluated at other data on social networks, including the number of tweets, geographic regions, and social connections on Twitter. To detect bullying, Chen et al. [12] built the social network architecture framework. Algaradi et al. [14] integrated a variety of information sources. To create a powerful detection model, they took into account the network, activity, user, and tweet content. Lu Cheng et al. [7] constructed a complicated heterogeneous network using metadata, such as an, video, user profile, time, location, and comments, to achieve better outcomes. By network embedding, they discovered the post's vector representation, after which they categorized the post using SVM, Random Forest, etc. These models make an effort to address the issue that the lack of text entails.

In addition to outlining the format of a bully episode and potential related roles, Xu, et al. [14] proposed various natural language processing algorithms to detect bully traces. Latent Dirichlet Analysis(LDA) was used to discover topics, and Sentiment Analysis was used to identify roles. As a binary (positive/negative) classification problem, cyberbullying detection is formulated, and a linear SVM is trained using manually labelled data. The results showed an 89% cross validation accuracy, demonstrating that even simple features and a common classifier can be helpful to identify textual indicators of cyberbullying.

Deep learning also has greater representation capabilities for text content when used as an end-to-end method. These techniques, which make use of convolutional neural networks CNN [16], Recurrent neural network (RNN) [17], and a combination of CNN and RNN (RCNN) [20], have largely improved text classification. Miswriting occurs frequently in social media, especially when bullying text is being sent. To address that, Park et al. [18] used a hybrid model. Convolutional neural networks were employed to classify at the character and word levels, respectively. Convolution layers and the Gate Recurrent Unit (GRU) were combined by Ziqi [21] to encode the text; the approach combined structural and sequence information. In order to pay closer to specific words, attention mechanisms are added in the cyberbullying detection process. To detect bullying language, Zhang [19] suggested a bidirectional RNN (BiRNN) [22] model with the attention mechanism. This area is usually the attention mechanism to adjust the words' weight while integrating the contextual feature from BiRNN. Similar to the methods mentioned above, the deep model has benefited from certain meta-data. Using a combination of the latent representation from text and metadata, Founta [15] trained the hybrid model.

The current

**Proposed System:**

Artificial intelligence has been improved tremendously without needing to change the underlying hardware infrastructure. Users can run an Artificial intelligence program in an old computer system. On the other hand, the beneficiary effect of machine learning is unlimited. Natural Language Processing is one of the branches of AI that allows machines to read, understand, and deliver meaning. NLP has been very successful in healthcare, media, finance, and human resource. Text classification is a common NLP task that assigns a label or class to text. There are many practical applications of text classification widely used in production by some of today’s largest companies. One of the most popular forms of text classification is sentiment analysis, which assigns a label like positive, negative, or neutral to a sequence of text.

Text data derived from natural language is unstructured and noisy. A natural language processing system for textual data reads, processes, analyzes, and interprets the text. As a first step, the system preprocesses the text into a more structured format using several different stages. Text preprocessing is an essential step in building a Machine Learning model and depends on how well the data has been preprocessed. Text preprocessing involves transforming text into a clean and consistent format that can then be fed into a model for further analysis and learning. Text preprocessing techniques may be general so that they apply to many types of applications, or they can be specialized for a specific task. For example, the methods for processing scientific documents with equations and other mathematical symbols can be quite different from those for dealing with user comments on social media. However, some steps, such as lowercase, sentence segmentation, tokenization, spelling corrections, and stemming, are common.

Transfer learning is a technique where a deep learning model trained on a large dataset is used to perform similar tasks on another dataset. We call such a deep learning model a pre-trained model. The most renowned examples of pre-trained models are the computer vision deep learning models trained on the ImageNet dataset. Similarly, we will use the pre-trained models on the natural language processing task which are trained on the large text corpus data and we can fine-tune the model on our dataset to get a better understanding of the semantic nature within the text language words. In the past two years, there have been significant improvements in using neural network–based text representations for NLP tasks. There are some of the “Universal Text Representations”. You can see why there’s been a surge in the popularity of pretrained models. We’ve seen the likes of Google’s BERT and OpenAI’s GPT-2 take the bull by the horns. These representations have been used successfully for text classification in the recent past by fine-tuning the pre-trained models to the given task and dataset. BERT is a popular model used in this way for text classification.

Vaswani introduced a new form of attention, self-attention, and with it a new class of models,. A Transformer still consists of the typical encoder-decoder setup but uses a novel new architecture for both. The encoder consists of 6 Layers with 2 sublayers each. The newly developed self-attention in the first sublayer allows a transformer model to process all input words at once and model the relationships between all words in a sentence. This allows transformers to model long-range dependencies in a sentence faster than RNN and CNN-based models. The speed improvement and the fact that individual attention heads learn to perform different tasks’’ lead to the eventual development of Bidirectional Encoder Representations from Transformers. BERT and its successors are, at the time of writing, state-of-the-art models used for transfer learning in NLP.

BERT (Bidirectional Encoder Representations from Transformers is published by researchers at Google AI Language group. It is regarded as a milestone in the NLP community by proposing a bidirectional Language model based on Transformer. BERT uses the Transformer Encoder as the structure of the pre-train model and addresses the unidirectional constraints by proposing new pre-training objectives: the “masked language model”(MLM) and a “next sentence prediction”(NSP) task. BERT advances state-of-the-art performance for eleven NLP tasks and its improved variants Albert and Roberta also reach great success. So further we are going to explore and fine-tune the state-of-the-art models like Albert, Roberta, et,. On our dataset to generate better accuracy with these language models.

As our dataset is an imbalanced dataset training the model is a challenging task with hyperparameters such as learning rate, the number of epochs, batch size, optimizers, sequence lengths, schedulers, etc., To handle the imbalanced dataset we have different techniques such as label weight balancing, oversample the under sampled data by creating or adding new data using the text data augmentation techniques. The dataset is split into train and test with equal proportions using the stratified split so that an equal number of class labels are split in to both the training and testing phases. The selection of the pre-trained models is another task where we are going to check with multiple pretrained language models and compare the models by evaluating the test data with finetuned hyperparameters and then the best model is selected to evaluate the test dataset.

**Implementation:**

Algorithm used:

Roberta Model:

RoBERTa stands for Robustly Optimized BERT Pre-training Approach. It was presented by researchers at Facebook and Washington University. The goal of this paper was to optimize the training of BERT architecture to take lesser time during pre-training. BERT is basically an Encoder stack of transformer architecture. A transformer architecture is an encoder-decoder network that uses self-attention on the encoder side and attention on the decoder side. BERTBASE has 12 layers in the Encoder stack while BERTLARGE has 24 layers in the Encoder stack. These are more than the Transformer architecture described in the original paper (6 encoder layers).

BERT architectures (BASE and LARGE) also have larger feedforward-networks (768 and 1024 hidden units respectively), and more attention heads (12 and 16 respectively) than the Transformer architecture suggested in the original paper. It contains 512 hidden units and 8 attention heads. BERTBASE contains 110M parameters while BERTLARGE has 340M parameters. BERT was able to improve the accuracy (or F1-score) on many Natural Language Processing and Language Modelling tasks. The main breakthrough that is provided by this paper is allowing the use of semi-supervised learning for many NLP task that allows transfer learning in NLP. It is also used in Google search, as of December 2019 it was used in 70 languages.

Chart, histogram

Description automatically generated

We have trained the Dataset from open source where we have 2 classes which are non-offensive and offensive. To get the accurate prediction, we need to discard the items which are less than 5 words in the statement. The above illustration depicts the frequency of documents of a given length with respective to density.

Chart, line chart

Description automatically generated

Here we used Roberta classifier to train the model and used 12 number of Epochs. Optimizer we used here is AdamW and the scheduler is get linear schedule with warm up. Splitted the trained data by 80%, Valid data by 10% and test data by 10%. When we trained 1 of 12 Epochs we got the train loss to be 0.5462 and valid loss to be 0.3399, 2 of 12 Epochs we got the train loss to be 0.2147 and valid loss to be 0.1456, 3 of 12 Epochs we got the train loss to be 0.1464 and valid loss to be 0.1318, 4 of 12 Epochs we got the train loss to be 0.1245 and valid loss to be 0.1243, 5 of 12 Epochs we got the train loss to be 0.1147 and valid loss to be 0.1218, 6 of 12 Epochs we got the train loss to be 0.1064 and valid loss to be 0.1194, 7 of 12 Epochs we got the train loss to be 0.0994 and valid loss to be 0.1206, 8 of 12 Epochs we got the train loss to be 0.0926 and valid loss to be 0.1235, 8 of 12 Epochs we got the train loss to be 0.0926 and valid loss to be 0.1235, 9 of 12 Epochs we got the train loss to be 0.0876 and valid loss to be 0.1221, 10 of 12 Epochs we got the train loss to be 0.0830 and valid loss to be 0.1177, 11 of 12 Epochs we got the train loss to be 0.0797 and valid loss to be 0.1173, Likewise when used 12 of 12 Epochs we got the train loss to be 0.0753 and valid loss to be 0.1130. As we can see that the train loss and valid loss is gradually decreasing just shown as in the graph above.

Chart, treemap chart

Description automatically generated

Here’s the confusion matrix when we train our model, we have taken 2065 non offensive statements out of which 65 statements gave us wrong predictions and 2000 statements got right predictions, Similarly we have trained 414 offensive statements in which 343 statements were predicted correctly and the rest 71 were wrong predictions. When we train our model, from 2065 non-offensive statements we got precision around 0.96%, recall with 0.96% and f1-score around 0.96%. Also from Offensive statements, we got precision around 0.84%, recall with 0.83% and f1-score was 0.83%. Therefore from RoBERTa model the average accuracy was around 0.94%.

Comparison Table:

|  |  |  |
| --- | --- | --- |
| Algorithms | Accuracy | Dataset |
| Logistic Regression | 76.8% | YouTube |
| Naïve Bayes | 60.6% | Reddit |
| Support vector machine | 64.8% | Wikipedia |
| J48 | 61% | Formspring |
| XGBoost | 77.4% | Twitter |
| Sequential minimal optimization | 68.47% | Twitter |
| Our Proposed method: RoBERTa model | 94% | Twitter |

**Results**

A picture containing scatter chart

Description automatically generated

Scatter chart

Description automatically generated

We can see that in above 2 pictures, when we type any word, it detects whether it comes under Offensive or Non-offensive statements. We took 40 tokens as the length of statement, when we enter any statement the starting token will be 0 and all the words in the statement will be converted into numbers and statement will be ended by token 2 by default.

**References:**

[1] Smith, Peter K.; Mahdavi, Jess; Carvalho, Manuel; Fisher, Sonja; Russell, Shanette; Tippett, Neil (2008). "Cyberbullying: its nature and impact in secondary school pupils".

[2] Waasdorp, T.E.; Bradshaw, C.P. (2015). "The overlap between cyberbullying and traditional bullying".

[3] Barlett, Christopher P.; Simmers, Matthew M.; Roth, Brendan; Gentile, Douglas (July 4, 2021), "Comparing cyberbullying prevalence and process before and during the COVID-19 pandemic".

[4] Smith, Alison M. (September 5, 2008). "Protection of Children Online: Federal and State Laws Addressing Cyberstalking, Cyberharassment, and Cyberbullying".

[5] Tereza Iofciu, Péter Fankhauser, +1 author Kerstin Bischoff, prieto et al. 2009. “Identifying Users Across Social Tagging Systems”.

[6] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, “How unique and traceable are usernames?” in Proc. 11th Int. Conf. Privacy Enhancing Technol., 2011, pp. 1–17.

[7] Mitchell, Tom (1997). Machine Learning. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892….

[8] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, “Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network,” Computers in Human Behavior, vol. 63, pp. 433–443, 2016.

[9] 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, 2015.

[10] “Us and them: identifying cyber hate on twitter across multiple protected characteristics,” EPJ Data Science, vol. 5, no. 1, p. 11, 2016.

[11] V. S. Chavan and S. Shylaja, “Machine learning approach for detection of cyber-aggressive comments by peers on social media network,” in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2015, pp. 2354–2358.

[12] C. Chelmis, D.-S. Zois, and M. Yao, “Mining patterns of cyberbullying on twitter,” in 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017, pp. 126–133.

[13] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, “Xbully: Cyberbullying detection within a multi-modal context,” in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM, 2019, pp. 339–347.

[14] Xu, Jun-Ming; Kwang-Sung Jun; Xiaojin Zhu; and Amy Bellmore. Learning from bullying traces in social media. In Proceedings of the NAACL HLT Conference, Montreal, Canada, 2012, pp.656-666.

[15] A.-M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, “A unified deep learning architecture for abuse detection,” arXiv preprint arXiv:1802.00385, 2018.

[16] B. Gamback and U. K. Sikdar, “Using convolutional neural networks to ¨ classify hate-speech,” in Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.

[17] H. Hosseinmardi, S. Li, Z. Yang, Q. Lv, R. I. Rafiq, R. Han, and S. Mishra, “A comparison of common users across instagram and ask. fm to better understand cyberbullying,” in 2014 IEEE Fourth International Conference on Big Data and Cloud Computing. IEEE, 2014, pp. 355– 362.

[18] J. H. Park and P. Fung, “One-step and two-step classification for abusive language detection on twitter,” arXiv preprint arXiv:1706.01206, 2017.

[19] A. Zhang, B. Li, S. Wan, and K. Wang, “Cyberbullying detection with birnn and attention mechanism,” in International Conference on Machine Learning and Intelligent Communications. Springer, 2019, pp. 623– 635.

[20] R. Zhang, H. Lee, and D. Radev, “Dependency sensitive convolutional neural networks for modeling sentences and documents,” arXiv preprint arXiv:1611.02361, 2016.

[21] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on twitter using a convolution-gru based deep neural network,” in European Semantic Web Conference. Springer, 2018, pp. 745–760.

[22] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, “Attentionbased bidirectional long short-term memory networks for relation classification,” in Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), 2016, pp. 207– 212.