IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9515

Automated detection and Risk Assessment of Cyberbulling by Implementation of a Comment Toxicity Detector

Apaar Gandhi, Dilip Kumar

Maharaja Agrasen Institute of Technology

Abstract - Deep mastering techniques have lately all started for use to stumble on abusive comments made on line boards. Detecting, and classifying online abusive language is a non-trivial NLP task due to the fact online remarks are made in a huge style of contexts and incorporate words from many distinctive formal and casual lexicons. moreover, spelling and grammar errors (many of them intentional) abound. In this paper, we observe and put into effect baseline and existing procedures for the project of classifying online abuse, andadditionally introduce and examine editions of the present fashions. Our goal is to offer a scientifically rigorous perspective on the strengths and weaknesses of the variety of processes. As such, we practice each methodto 2 extraordinary facts sets and offer in-depth visualizations of model performance and explanatory wins and losses.

Keywords: Toxic Comments, Natural Language Processing, Machine Learning, Deep Learning, Text Classification, Multilabel Classification

I. INTRODUCTION

The advances in IT technology and generalizing virtualization all over the globe have caused unheard-of participation in social media, and there's absolute confidence that social media is one of the biggest hallmarks of the twenty-first century but behind the defense of computer systems as digital partitions, some people additionally suppose they can abuse and harass different people's reviewsand characters, which may be coined as the period 'Cyber bullying'. This includes commonplace 'cyber bullying' procedures consisting of threatening to harm someone, posting suggests hurtful, or embarrassing comments, and callcalling. things related to this type of online harassment lead to a vital region of facts science to be capable of separating and distinguishing harassment feedback and cyberbullying, referred to as toxic feedback, from ordinary comments.

A studies initiative founded by way of Jigsaw and Google is currently working on tools to assist enhance online conversations. One good-sized thing in their efforts is aimingto identify the toxic remarks and lunch an online toxicity tracking gadget on numerous online social structures. Over those years, we've seen a lot of instances where social media have performed a pivot position because of poisonous remarks and hatred. For an instance, the leader Minister of the Uttar Pradesh nation of India blamed social media like Facebook, Twitter, and YouTube for escalating tensions in the course of the communal conflict between the Hindu and Muslim communities in Muzaffar Nagar, India in 2013 [2]. Kalamboli police booked a man for abusing and dangerous to the police through a comment on Facebook put up [3]. another example is of Riots that occurred in DJ Hali, Bengaluru, India in 2020 over a provocative Facebook post against Islam that left 3 lifeless and plenty injured [4].

On January 6, 2021, the US Capitol Riots came about by way of a supporter of Donald Trump. Many extremists had published on Social Networking websites posts such as "occupy the Capitol", "carry revolution", and many others. before riots [5]. consequently, it's miles very critical to discover such threats, hatred, and toxicity on online dialogue systems and social networking websites. due to the fact no longer doing so can purpose violence, and riots, prevent exact debates, make the net a hazardous region, and can have an effect on human beings mentally.

In a joint effort with Kaggle, they described the task as a competition poisonous comment category challenge. thinking about and extending the problem announcement, the intention of the venture is to classify feedback into poisonous and now not toxic, and if poisonous then to which subcategory it belongs to. we can also further enlarge it to examine conventional machine learning algorithms with deeplearning algorithms, for this reason, aiming to discover and expand an efficient algorithm to become aware of poisonous feedback. a good way to achieve our goal on this take a look at, related remarks records of Crowdsourcing (159,000 textual content feedback) that was previously classified as seven classes had been used.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9515

We will use distinct system mastering and deep getting to know models on our data set that are made available by using traditional AI in Kaggle.com. we will use Logistic Regression and aid Vector system models with TF-IDF Vectorizer, lengthy brief-time period memory with Glove and Word2Vec Embedding. we've got used all on the given dataset and examine their ratings to locate which one may beexcellent.

II. RELATED WORK

Its associated studies have looked into hate speech, online harassment, abusive language, cyberbullying, and offensive language. commonly talking, poisonous comment detection is a supervised category undertaking and may be approached through both manual feature

engineering and neural networks. Nguyen [12] made a model including 2 additives deep getting to know Classifier and Tweet Processor. Tweet Processor is used for making use of semantic regulations and preprocessing datasets to capture essential information. Their model produces an accuracy of 86. sixty-three% on Stand ford Twitter Sentiment Corpus. numerous studies have formally investigated hate speech using neural community strategies; Badjatiya et al. used large experiments with multiple deep gaining knowledge of architectures to analyze semantic word embedding to address poisonous comments identity [13]. In some other examine, sentiment analysis model of YouTube video feedback, the usage of adeep neural community become proposed that led to 70- eighty% accuracy [14]. Hossein Hosseini et al. [15] followthe assault on the perspective toxic comment detection website. This internet site offers a poisonous rating to any word. They tried to adjust a poisonous phrase having equal that means so that model will supply it with very low toxic rankings. This life is dangerous for poisonous detection gadgets. also, popular use of different forms of neural community techniques for remark category had been drastically utilized in the currently published literature [6,16,17]; however, these approaches simplest addressed anumber of the assignment's demanding situations whilst others continue to be unsolved. furthermore, Farag and El- Seoud[18] suggested that full-size numbers of literature have shown that supervised getting to know techniques had been the maximum often used methods for cyber- bullying detection. nevertheless, other non-supervised strategies and methods have been diagnosed to be operative in cyber-bullying popularity. additionally, Karlekar and Bansal [19] mentioned an elevated number of private sexual harassment and abuse which are shared and published online. authors presented the challenge of robotically categorizing and studying various styles of sexual harassment, primarily based on testimonies shared on the web discussion board safe city and used labeling degrees of groping, ogling, and commenting; their results indicated that the single-label CNN-RNN model achieves an accuracy of 86. five. one of the primary undiscovered problems is the way to discover algorithms that are capable of put in force high sensitivity in the detection of toxic comments. Of direction, figuring out a comment that is not poisonous as poisonous may be frustrating for the users and there ought to be a lot of attempts to form a set of rules with the highest diploma of sensitivities. Jigsaw and Googles Counter Abuse technology team introduce a task named perspective. It uses devices to gain knowledge of fashions to become aware of abusive remarks. The fashions score a phrase-based totally on the perceived effect the textual content may have in a conversation and have the functionality to classify comments. Navone Chakrabarty [18] makes use of the machine getting to know model on the Jigsaw poisonous remark Detection dataset to label toxicity of remarks and bring the implied Validation Accuracy, so received, is 98%.

III. BASIC DESCRIPTION OF THE SOLUTION

In this paper, we took a dataset provided by the Kaggle website provided by using traditional AI. it's a far collection of a large number of remarks on the Wikipedia website and categorized as – poisonous, extremely toxic, risk, identity hate, obscene, and insult. The advantage of this type of data is that these remarks represent a real pattern of the content present on social media sites. We first ran an analysis and visualization on this fact which we've mentioned in phase III-B. For our gadget mastering model, we've eliminated outliers and noise that is present in information. We start with tested the overall performance of classical fashions particularly, guide Vector gadgets and Logistic Regression on this challenge. We then carried out pre-trained embeddings, specifically Glove and Word2Vec in our version and carried out the classification.

A. Type of classification

As mentioned above our dataset have 6 classes i.e., hazard, insult, toxic, severe toxic, obscene, or identification hate. consequently our hassle can belong to multiclass or multi- label class hassle. As we are able to see the above description this hassle is a multiclass classification as well as multilabel type hassle.

- Multiclass classification: A multiclass type is a machine gaining knowledge of a classification mission that consists of greater than two lessons, or outputs. as an example, the use of a model to identify animal types in pics from an encyclopedia is a multiclass category instance because there are many extraordinary animal classifications that every photo may be classified.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9515

- Multilabel class: Multi-label type is a generalization of multiclass category, which is the single-label hassle of categorizing times into exactly certainly one of greater than instructions.

B. Metrics

To be capable of evaluating the overall performance of each set of rules, several. There are quite a several evaluation metrics for device gaining knowledge of models. The problem includes an incredibly unbalanced dataset. So, accuracy isn't a nicely-applicable performance degree. With the best 10% of the schooling information belonging to the nice magnificence (hate tags), it's far trivial to attain ninety% accuracy via a naïve model which labels each entry as easy. Precision-take into account or F1 score look like the subsequent obvious preference however they have their proportion of boundaries which includes choice of the threshold value and relative significance to receive to precision vs bear in mind. for this reason, we finally settled at the curve and score which provide a completely accurate photo of the performance of a discriminative model, Hamming loss and Log loss.
The receiver working feature is a curve that plots properpositive rate (TPR) vs false tremendous fee (FPR).

$$TPR = TP / TP + FN$$

where TP (True Positive) is several samples that are true and predicted as true and FN (False Negative) is several samples that are false and predicted as true.

$$FPR = FP / FP + TN$$

in which, FP (false fine) is a wide variety of samples that are false and anticipated as fake and TN (real bad) is a wide variety of samples that can be proper and anticipated as fake.
AUC denotes the whole region under the ROC curve for thegiven area. Its cost can range from 0 to one. A better version will have a greater location underneath the ROC curve with a great model having AUC=1 and a model which usually predicts incorrectly having AUC score=zero. in this sense AUC may be understood because the common of performance measures of the classifier across all thresholds. AUC is scale and threshold invariant.
Hamming Loss is a fraction number of labels that areincorrectly predicted to a general number of labels.

$Hamming\ Loss = 1/\ NL\ \sum\sum Yi,l \oplus Xi,l\ N\ i{=}1\ L\ l{=}1$ where, $\oplus$ is exclusive-or, Yi,l is the predicted value, and Xi,l is the actual value for the ith comment on lth label
value, NL is the total number of labels.
Log Loss takes into account the probability of models. It isdefined as the following:

$Log\ Loss = 1/N\ \sum\sum Yli \log(pli)\ M\ i{=}1\ N\ l{=}1$ where M is the number of labels, N is the number of samples, Yli is a binary indicator of the correct classification
and is model probability.

IV. DATA PREPARATION

One of the primary challenges in class instances is having appropriately labeled information, wherein a representative training set will be extracted for modeling. For textual content-based totally or natural Language Processing [NLP] problems, this predicament is even extra pronounced [24]. asthe sentiment inference of written verbal exchange is very subjective, there's little preference but to leverage human labeling instead of a proper analytical labeling model (consequently the need for the classification algorithm). eventhough many big unlabeled textual content corpora are easy to be had, human classified data is lots greater rare [25]. luckily, for diverse altruistic reasons, several businesses are presenting open-sourced categorized statistics sets1. One such source is the Wiki Media Foundation2 which offers get admission to textual content comment snippets from Wikipedia to speak pages. Crowd-sourcing become used to manually label over 159,000 text remarks, flagging each as one in all seven following alternatives. we've used both gadget studying and Deep gaining knowledge of fashions over our dataset. we've used consequently used TF-IDF Vectorizer with device mastering fashions – Logistic Regression and SVM and Glove and Word2Vec Embedding with LSTM. the uncooked textual content can not directly be

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9515

input to any machine gaining knowledge of algorithms. someof the maximum famous techniques for changing textual content into numerical capabilities are Bag of words, TF-IDFWord2Vec, and Glove.

1. Term Frequency Inverse Document Frequency (TF- IDF)

Vectorizer - vectorizer normalizes the textual content. Itreduces the weight of tokens which can be going on a coupleof instances in files. it's far made in this type of manner that its cost will increase if the frequency of period is more in a report and value decreases if term happens in multiple files.

It consists of 2 parameters – Term Frequency (TF) andinverse report frequency (IDF).

- Term Frequency–$TF(i, d)$: This calculates the frequency of a time $'i'$ in a file 'd'. this is just like the count Vectorizer approach to encoding text.

- Inverse document Frequency-$IDF(t)$: IDF gives the inverse of record frequency. dft relies on documents that carry a time t. therefore, it calculates several files in which term seems and takes inverse and logs that. Later 1 is the denominator delivered to avoid 0-division.

$idf(t)$=log (1+ $|D|$ 1+ dft)+ 1

Wherein, dft denotes the number of files containing time t and $|D|$ includes the total range of files.

2. Glove Word Embedding

Glove word embeddings are used to represent words in an established layout. because the maximum of the records on the net isn't based, word embeddings techniques are a usefultool to convert facts into the extra dependent layout so that useful facts can be extracted.
In Bag of words, models feature extraction may be accomplished but they fail to capture any semantic or contextual records in texts. One-warm encoded vectors may also lead to a tremendously sparse shape which causes the version to overfit. to triumph over these shortcomings of theabove approach, word embeddings are used.
Word embeddings constitute phrases in the shape of vectorsin a pre-described dense vector area. those vectors contain meaningful semantic facts approximately the words. The idea of this method is phrased with similar semantic data like each other. subsequently, similar words can be in proximity in the high dimensional vector area. hence, we can drastically lessen the vector size in contrast to the one- encoding method.
Pretrained phrase embeddings are received by way of unsupervised education of a model on a massive corpus. As they're trained on a huge corpus, they seize the semantic data of most of the phrases. those pretrained embeddings are provided by way of distinct organizations and businesses foropen use.
Glove stands for worldwide Vectors. it's miles furnished by Stanford as an open-supply challenge. in this method, a word co-occurrence matrix is constructed. This allows for shooting the semantic statistics. The co-occurrence matrix

stores record the frequency that seems in some context. as a result, it considers each local record and international statistics to reap the embeddings.

3. Word2Vec

It's far one of the earliest pretrained embeddings. It has 2 flavors. First is a skip-Gram model where the set of rules tries to are expecting the context of surrounding words in which the word could be used. It learns by way of predictingthe surrounding phrases given a contemporary word. 2nd is the continuous Bag of phrases (CBOW) version where the set of rules attempts to predict the phrase if a context is given. In this way, the word embeddings vectors are generated.

D. Machine Learning Models

1. Logistic Regression

Logistic Regression is the proper regression evaluation to apply while the reliant variable has a binary answer. like all closing varieties of regression frameworks, Logistic Regression is likewise a predictive regression framework. Logistic Regression is utilized to evaluate the relationship between one reliant variable and one or greater non-reliant variable. It gives discrete yields going someplace within therange of 0 and 1. Logistic Regression utilizes a greater complicated fee feature; this price function is called the 'Sigmoid feature' or in any other case called the 'logistic feature'.
$f(x)$ = 1/1+ e^{-x}
In our case, we've got used logistic regression for prediction in every magnificence i.e., toxic, extremely poisonous, chance, identification hate, obscene and insult and mean rating.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9515

2. Support Vector Machine

Vector Machines use the idea of support vectors and hyperplanes for class obligations. They can be used for each regression and class obligation but are more famous for classification. They can be used for each linear and non- linear fact. it works by constructing a foremost hyperplane in an N-dimensional space i.e., a boundary isolating the facts points, such that the margins between the help vectors are maximized. help vectors refer back to the records factors which are closest to the hyperplane and are useful for training tasks, and the margin refers to the gap between the two parallel traces passing via the closest guide vectors on either facet of the hyperplane. In the case of non-linearly separable records, it transforms the unique facts into better dimensions for category mission.

it is also known that SVM carries out remarkable for higher dimensional information because the complexity of SVM does now not rely on the dimensionality of information usedbut on the number of guide vectors. This additionally helps it to be memory efficient. support Vector system In our case, we've got used the support Vector system with Binary Relevance and Classifier Chains for predictions. In the Binary Relevance technique, we transform the hassle into separate unmarried-class category issues, every one of theproblems having an unmarried label.

We then apply the assist Vector system to every problem one at a time to get the result. After that, the outcomes of each issue may be mixed to get all the labels for a remark. simple it's miles it comes with drawbacks. It ignores any correlation between the labels. therefore it will provide negative results if there may be a correlation among labels.In the Classifier Chain method, we transform the trouble into separate unmarried-label class issues, such that if the

ith classifier is trained on the input variable(s) X, then $(i + 1)th$ classifier is educated on input variable X and output produced with the aid of ith classifier. therefore, this method considers the correlation between the labels, because, for each new classifier, the predictions of the previous classifiers are taken into consideration, i.e., for a given target variable, it also considers the correlation amongprevious goal variables.

3. Long Short Term Memory

An artificial neural community is a layered layout of related neurons, enlivened by using a natural neural network. It isn't one set of rules yet mixes different algorithms which permit us to do complicated manner on facts. Recurrent Neural Networks (RNN) is a category of neural networks custom- designed to manage worldly data.

The neurons of RNN have a cell country/memory, and entry is handled via this interior kingdom, that's executed with thehelp of loops inside the neural networks. there may be a repeating module of '$tanh$' layers in RNNs that permit them to preserve statistics however not too lengthy. that is the purpose we need LSTM fashions. LSTM is a unique recurrent neural network that may seize long-term dependencies.

The mobile kingdom has regulated the usage of gates which decide the number of facts and a good way to drift through them. It has a cell kingdom ct alongside hidden states that store statistics. These statistics can travel through the cell kingdom with no alternative, therefore, maintaining long- time dependencies. determine 7 LSTM [19] LSTM unit takes modern-day enter, preceding hidden kingdom, preceding mobile country as the enter and effects within the new mobile country and hidden state.

An LSTM unit includes enter Gate, forget about Gate, OutputGate, Candidate Layer, Output Layer. all of the gates make use of the Sigmoid characteristic as an activation feature. overlook Gate Layer decides on the records to be saved in cell state the use of $ht−1$, xt and sigmoid layer. The decision of new data to be saved is executed by entering the gate layer and $tanh$ layer. $tanh$ creates a vector of candidates $\tilde{Ct}$ that may be new information. This takes place after the input gate decides on which values to replace. Output is decided on the foundation of the cellular state. The sigmoid layer decides which part of the cellular state to output. $tanh$ changes the price of cell kingdom in −1 and 1 and multiplies it by way ofthe output of the sigmoid gate.

RNN and LSTM give output based totally on modern statistics and beyond statistics that have already been surpassed via it. One directional LSTM does now not soak up account information similarly in sequence whilst predicting. Bi-directional LSTM trains two impartial LSTMs in oppositeinstructions and joins both the hidden layers to the equal utput. One LSTM trains in the forward route and the other in the backward path. With the use of the 2 hidden states mixed we can use facts from both the past and future.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9515

Model	Hamming Loss	Log Loss	Accuracy
Logistic Regression model	2.585258 5852585	1.8439563 996112658	0.89738973 89738974
SVM binary Relevance	2.9102910 291029103	1.49364645 6884211	0.8979897 98979898
SVM ClassifierChains	2.8352835 283528353	1.2667634 664895946	0.90069006 90069007
LSTM (Word2VecEmbedding)	0.02778819 802640491	0.41639971 73309326	0.94261944 29397583
LSTM (Glove Embedding)	0.02574582 929548699	0.33832901 7162323	0.97549253 7021637

In the prediction of the next phrase problem, Unidirectional

After evaluating the consequences, we can say that LSTM with Glove embedding plays the best as it has maximum accuracy count and least Hamming loss and one of the leastLog loss among all fashions which means that there may be very much less multilabel are appropriately measured.

LSTM with Word2Vec embedding has also accomplished corresponding to LSTM with glove embedding. We also look at that Classifier SVM performs higher than Binary Relevance SVM which become expected. each hamming loss and log loss in all our fashions are decreased than the algorithms supplied in [14]. It changed into expected for deep gaining knowledge of version LSTM to have the bestresult overall the algorithms.

LSTM can best see “The female went to …” but in Bidirectional LSTM, ahead LSTM sees “The girl went to …”and a backward LSTM sees “… and then there has been sandstorm”. These statistics provided using Backward LSTM can assist to understand what the subsequent phrase is. For our case, we've used Word2Vec and Glove word embeddingto be had in Kaggle with 300 dimensions after which teach aBidirectional LSTM with four epochs.

The facts set are offered through a closed Kaggle competition. each file inside the facts set is a textual consumer comment with six binary fields for every of the sixtoxicity ranges, with null being non-toxic. The initial task become to reduce the seven labels down to two (toxic, non- toxic) classes. This generalization allowed us to awareness ofa greater correct binary classifier.

After finding and imparting information for this have a look, preprocessing and cleansing the data were the first boundaries that challenged this project. it's far very critical to any NLP version in real international demanding situations. for example, ‘don't’, ‘do no longer, and ’don’t’ have all similar which means that without pre-processing they're considered specific tokens. The layout of the pre-processing section is certainly undertaking specific. in the modern-day examination, we assume that the verb tenses or similar wordshad been not important for being identified. consequently, lemmatization was applied to reduce vocabulary to four kindsof nouns, verbs, adjectives, and adverbs. the use of the very simple intuitive strategies in our architectures had been quite useful in NLP duties. no matter the final performance, they offer the right instinct approximately the task..

V. RESULT

After applying 3 different machine learning models to ourdataset, we got the result in form of Accuracy, HammingLoss and Log Loss.

VI. CONCLUSION

We compared the overall performance of the version based on the suggested accuracy ratings, hamming, and log loss. Hamming and Log's lack of classical fashions are extra than deep gaining knowledge of the LSTM version. We additionally determined that the classifier chain method finished slightly higher than binary relevance on this venture. LSTM model outperformed different models on this task. We can also experiment with more sophisticated fashions like GRUs. We can also combine gadget studying models for this trouble.

REFERENCES

[1] "Toxic Comment Classification" by Sara Zaheri, Jeff Leath et al. (smu.edu) (April 2020)

[2] B. S., "The Role of Social Media in Mobilizing People for Riotsand Revolutions," in Social Media in Politics,

DOI: 10.17148/IARJSET.2022.9515

Public Administration and Information Technology, vol 13. https://doi.org/10.1007/978-3-319-04666-2_19, Springer, Cham , 2014.

[3] R. Assainar, "The Hindu," 04 May 2020. [Online]. Available: https://www.thehindu.com/news/cities/mumbai/kalamboli- manabuses-police-on-fb-booked/article31496732.ece.

[4] A. Bharadwaj, "The Hindu," 12 August 2020. [Online].Available:https://www.thehindu.com/news/national/a t-least-three-killed-inpolice-firing-as-riots-break-out-over-fb- post inbengaluru/article32331790.ece.

[5] K. Dilanian, "NBC News," 8 March 2021. [Online]. Available: https://www.nbcnews.com/politics/justice-department/fbi- officialtold-congress-bureau-can-t-monitor-americans-social- n1259769.

[6] *Estela Saquete, David Tomas, Paloma Moreda, Patricio Martinez-Barco, and Manuel Palomar. Fighting post-truth using natural language processing: A review and open challenges. Expert Systems with Applications, 141:112943,2020.*

[7] *Alexandre Ashade Lassance Cunha, Melissa Carvalho Costa, and Marco Aur´elio C Pacheco. Sentiment analysis of youtube video comments using deep neural networks. In International Conference on Artificial Intelligence and Soft Computing, pages 561–570. Springer, 2019.*

[8] Alisha de Bruijn, Vesa Muhonen, Tommaso Albinonistraat, Wan Fokkink, Peter Bloem, and Business Analytics. Detecting offensive language using transfer learning. 2019.

[9] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pages5754–5764, 2019.

[10] *Nadine Farag, Samir Abou El-Seoud, Gerard McKee, and Ghada Hassan. Bullying hurts: A survey on non-supervised techniques for cyber-bullying detection. In Proceedings of the 2019 8th International Conference on Software and Information Engineering, pages 85–90, 2019.*

[11] *Keita Kurita, Anna Belova, and Antonios Anastasopoulos. Towards robust toxic content classification. arXiv preprint arXiv:1912.06872, 2019.Thomas A Birkland. An introduction to the policy process: Theories, concepts, and models of public policy making. Routledge, 2019.*

[12] *Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760, 2017.*

[13] *H. K. J. H. G. S. Rahul, "Classification of Online Toxic Comments Using Machine Learning Algorithms," in Proceedings of the International Conference on IntelligentComputing and Control Systems , 2020.*