Systematic Literature Review: Toxic Comment Classification

2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA) | 978-1-6654-6136-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICITDA55840.2022.9971338

Felix Museng
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
felix.museng@binus.ac.id

Adelia Jessica
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
adelia.jessica@binus.ac.id

Nicole Wijaya
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
nicole002@binus.ac.id

Anderies Anderies
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
anderies@binus.ac.id

Irene Anindaputri Iswanto
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
irene.iswanto@binus.ac.id

***Abstract*—Over the last decade, deep learning models have surpassed machine learning models in text classification. However, with the continuity of the digital age, many are exposed to the dangers of the internet. One of the dangers would be cyberbullying. In an attempt to decrease cyberbullying, much toxic text detection and classification research has been done. In this paper, we aim to understand the effectiveness of deep learning models compared to machine learning models along with the most common models used by researchers in the last 5 years. We will also be providing insight on the most common data sets utilized by researchers to detect toxic comments. To achieve this, we have compiled the datasets of research papers and analyze the algorithm used. The findings indicate that Long Term Short Memory is the most routinely mentioned deep learning model with 8 out of 26 research papers. LSTM has also repeatedly yielded high accuracy results with above 79% for around 9000 data which could be adjusted depending on the pre-processing method used. There have been attempts to combine more than one deep learning algorithms, however these hybrid models might not result in a better accuracy than an original model. Furthermore, the most frequent sources of datasets came from Kaggle and Wikipedia datasets and a total of 13 researchers that used Wikipedia's talk page edits as their dataset.**

***Keywords*—Deep Learning, Machine Learning, Toxic Text Classification, LSTM**

I. INTRODUCTION

Toxic is a term commonly used to describe something as unpleasant. Toxic comments are comments made by users that use unpleasant words which could offend other users. These unpleasant words could be rude, disrespectful, and degrading.

With the increasing trend of social media, cyberbullying has quickly become one of the most concerning issues in this modern era. More often than not, a person would most likely have had experience of being cyberbullied or even witness a cyberbullying case. Although there are some unsevere cyberbullying, for example calling someone stupid online, there are also a lot of extreme cases of cyberbullying. Therefore, as a way to prevent cyberbullying, most companies have adopted some kind of toxic comment detection such as blurring or changing the toxic words into other symbols such as *. Currently, there are many methods used to detect toxic comments. Most of them are Machine Learning methods.

Machine learning has become one of the ways to detect toxic comments. Different machine learning algorithms were being used to classify toxic comments i.e., logistic regression, random forest, SVM (Support Vector Machine), Naive Bayes, decision tree, and KNN (K-Nearest Neighbors) [18]. Most researchers have done these algorithms in different datasets, but based on the accuracy score, some models still need more improvements.

Therefore, traditional machine learning methods tend to have a low precision for toxic comments detection as the variation of cyberbullying style is high and mostly focused on the usage of swear words, hence the usage of deep learning methods is used to improve the accuracy for detection of cyberbullying [19].

Deep learning is known as a part or subset of machine learning. Deep learning eliminates some data pre-processing and can distinguish important data that would alter the results. An example of deep learning would be being able to do things that came naturally as a human. A popular representation of deep learning implementation in the real world would be driverless cars. There are a couple of frequently mentioned algorithms in deep learning. Those are CNN (Convolutional Neural Networks) [10], RNN (Recurrent Neural Networks) [2], and LSTM (Long Short-Term Memory Networks) [9].

Deep learning has some advantages compared to machine learning. One of the advantages of deep learning is that it can handle a large amount of data to be processed, especially when it comes to complicated issues like picture classification, natural language processing, and speech recognition. The objective of this paper is to find relevant information from research papers on the toxic comment classification and systematically listing and comparing the existing research which then the information is used to help direct future research.

The research questions and the standards for the paper that we employed are included in section 2. We aim to understand how successful deep learning is in comparison to machine learning techniques. As a result, in section 3, we classify the datasets utilized and the algorithms employed in papers. Finally, we provide our paper's conclusion in section 4.

II. Study Review

A. Planning

The first step in the planning process is to identify the requirements for a particular systematic review. In the Introduction section, we discussed the necessity for a comprehensive review of machine learning algorithms for toxic comment classification. Following that, we established the following primary research questions:

- RQ1: Which data sets are utilized to detect toxic comments?
- RQ2: What are the most common deep learning models used to detect toxic comments?

Based on the research questions of the study, we have defined several electronic databases as the source of papers and journals. To ensure the eligibility of the references selected, Selection and Exclusion criteria are applied.

Table I. *Selection Criteria*

Inclusion Criteria	**Exclusion Criteria**
The paper must be related to classifying toxic comments with machine learning.	The work was published before 2018.
The paper must propose a method for filtering the toxic comments.	Non-relative studies to the research questions.
The work is recognized internationally.	-

B. Literature Review

In today's world, everyone has their own social media account where they may easily engage with others over the internet. This, however, can lead to cyberbullying, in which people try to make fun of others or harass, threaten, embarrass, or target someone else. For some people, this online brawl frequently turns into real-life threats, such as suicide [25].

Researchers are utilizing machine learning to detect and classify toxic comments to minimize cyberbullying. There are many models in machine learning to do it, such as Naive Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree, and K-Nearest Neighbor (KNN) [18].

However, in recent years, researchers prefer to use deep learning algorithms instead of machine learning. In the research [22], it was discovered that the proposed method, which combines CNN+LSTM, performs the best with The Unhealthy Comments Corpus (UCC) dataset available in GitHub which yields an accuracy score of 88.76%. It exceeded logistic regression and SVM, which had accuracy scores of 57.54 percent and 69.15 percent, respectively.

In another research [2], it is found that Multi-Emb RNN-CNN (0.59) outperformed Linear Regression (0.38) in F1 scores, further solidifying that Deep Learning algorithms perform better than Machine Learning algorithms. Taking reliability as a context, LIME model interpretability techniques could be used to analyze each method to help find the best model even if the scores are high [3].

Studies by Almerekhi et al [7], [6] utilize LSTM with GloVe on Reddit datasets to find variables that can help detect toxic comments such as toxic triggers, toxic posting behavior, and moderator rules strictness, the study also found evidence that some community is more robust against toxicity.

A comparative study by Zhao et al. [26] found the usage of language models such as BERT and RoBerta combined with CNN and BiLSTM on 4 different datasets. Another study by Kim et al. [6] utilizes GloVe, Word2Vec, Paragram, and FastText combined with BiLSTM, after implementing the topic-enhanced word embedding, the model is able to achieve an F1 score of 0.667.

C. Overview

The main research papers that we used were published in the range of 2018 until 2022 (within 5 years). The highest number of studies was in 2021. In 2022, there are just a few of them because it's the current year. In addition, all the research papers that we reference consist of fourteen international conference and journal papers, and twelve book chapters.

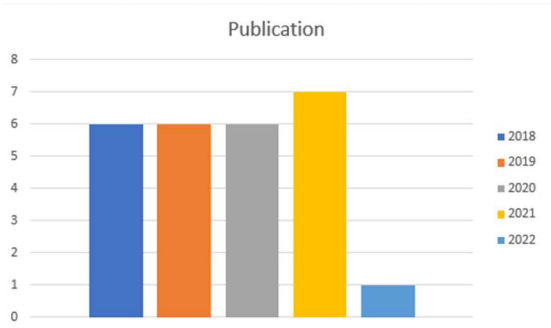

Fig 1. *Publication Year*

For the search criteria, we have given a detailed explanation of it in the table below (Table II).

Table II. *Search Criteria*

Database Searched	Search Terms / Keywords	Limits	Initial Search	Included in paper
IEEE Xplore	"toxic comment classification"	Between 2018 until 2022	35	[9], [11], [12], [18]
	"cyberbullying in social media"	Between 2018 until 2022, Publication topic: "learning", "deep learning"	49	[13], [25]
	"sentiment analysis deep learning"	Between 2018 until 2022,	758	[5]
Science Direct	"classify toxic comments deep learning"	Between 2018 until 2022, computer science subject areas, research articles	69	[1], [22]

(Continued) Table II. *Search Criteria*

Database Searched	Search Terms / Keywords	Limits	Initial Search	Included in paper
Springer	"toxic comment classification with machine learning"	Between 2018 until 2022, English, Conference Paper,	102	[15]
	"toxic comment classification"		143	[3]
	"detecting cyberbullying using deep learning"		151	[15]
IJTSRD	"toxic comment classification"	-	-	[16]
IARJSET	"toxic comment classification"	-	-	[17]
MDPI	"toxic comment classification"	Between 2018 until 2022	4	[5]
AIP Conference Proceedings	"toxic comment classification"	Between 2018 until 2022	8	[10]
ACM	"deep learning toxic comment classification"	Between 2018 until 2022	92383	[23], [2], [26], [20], [7], [6], [8], [4],
Semantic Scholar	"toxic comment classification"	Between 2018 until 2022, Type Journal Article	2810	[21] (published in KDIR Conference)
SMU data science review	"toxic comment classification"	Between 2018 until 2022	4	[24]

III. RESULT & DISCUSSION

A. RQ1: Which data sets are utilized to detect toxic comments?

In this section, we are aiming to learn more about the kind of data sets that are utilized in the research paper. To obtain a better understanding, we have created a list of tables (Table III) that is organized by the year of release. We divided the table into different columns, such as, the name of the datasets, total samples, dataset sources and the output distribution.

Table III. *Details of the datasets and its output distribution*

Reference	Datasets	Samples	Source	Output Distribution
2018 [12], [23], [2], [10], [17]	Wikipedia's talk page edits.	159571	Kaggle	6 classifications (toxic, severe toxic, obscene, threat, insult, and identity hate).
2018 [19]	Collaborative Knowledge Repository	100000	Wikipedia	3 classifications (personal attack, racism, and sexism).
	Microblogging	16000	Twitter	
	Q&A forum	12000	Formspring	
2019 [11], [16], [15], [21]	Wikipedia's talk page edits.	159571	Kaggle	6 classifications (toxic, severe toxic, obscene, threat, insult, and identity hate).
2019 [8]	WikiDetox	160000	Wikipedia	2 classifications (toxic, non-toxic)
2019 [4]	Quora insincere Question Classification Challenge	1306122	Kaggle	2 classifications (sincere, insincere)
2020 [9]	Twitter Comments	56745	Github	2 classifications (toxic, non-toxic)
2020 [18]	Online Toxic Comments	95981	Kaggle	6 classifications (threat, insult, toxic, severe toxic, obscene, or identity hate).
2020 [14]	Jingdong Platform Octopus Collector	14831	Jingdong	3 classifications (positive, negative, medium)
2020 [7],[6]	Reddit Comments on r/AskReddit Figure eight crowdsourcing platform	10000	Reddit	2 classifications (toxic, non-toxic)
2020 [24]	Wikipedia talk pages	159571	Kaggle	2 classifications (toxic, non-toxic)

(Continued) Table III. *Details of the datasets and its output distribution*

Reference	Datasets	Samples	Source	Output Distribution
2021 [1]	IMDB Review	50000	IMDB	2 classifications (positive, negative)
	COVID19-Fake	10700	COVID-19 Website	2 classifications (positive, negative)
2021 [13]	Wikipedia	3000	Wikipedia	1 classification (attack)
	Twitter	3000	Twitter	2 classifications (racism, sexism)
	Formspring	3000	Formspring	1 classification (bully)
2021 [26]	Wikipedia talk pages	159571	Kaggle	6 classifications (toxic, severe toxic, obscene, threat, insult, identity hate)
	Twitter Stream API (Founta)	50425	Twitter	4 classifications (abusive, hateful, normal, spam)
	Twitter and Corpus (Waseem)	18625	Twitter, Corpus	4 classifications (racism, sexism, both, neither)
	Facebook Comments (Kumar)	14998	Waseem	3 classifications (non-aggressive, overtly aggressive, covertly aggressive)
2021 [20]	Wikipedia	143000	Wikipedia	6 classifications (safe comments, toxic, obscene, threat, insult, identity hate)
	Youtube Comments	3700000	Youtube	6 classifications (safe comments, toxic, obscene, threat, insult, identity hate)
2021 [5]	Wikipedia talk pages	159,571	Kaggle	6 classifications (toxic, severe toxic, obscene, threat, insult, identity hate)
2021 [3]	Jigsaw Toxic Comment Classification Challenge / Wikipedia talk pages	159,571	Kaggle	6 classifications (toxic, severe toxic, obscene, threat, insult, identity hate
	Hate Speech Dataset	N/A	Github	2 classifications (toxic, non-toxic)
	Sexual Abusive YouTube Comments	1000	Zenodo	2 classifications (toxic, non-toxic)
2022 [22]	The Unhealthy Comments Corpus (UCC)	44355	Github	8 classifications (antagonize, condescending, dismissive, generalization, generalization unfair, healthy, hostile, and sarcastic).

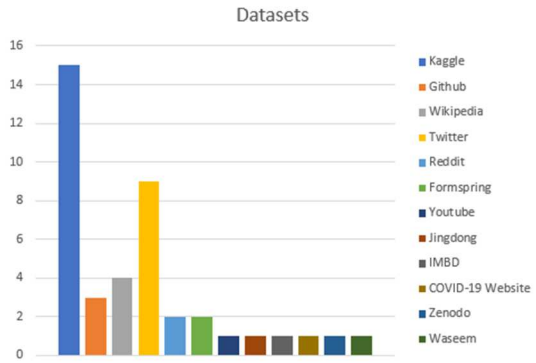

Fig 2. *Datasets used*

As shown in Fig 2, datasets that are being used from our references are mostly Kaggle, Twitter and Wikipedia. In a five years period, Wikipedia's talk page edits in Kaggle were the most popular datasets from five different researches. These top three sources are often used because of the large selection of datasets that can be used to help analyze the effectiveness of their algorithms in order to classify toxic comments.

B. RQ2: What are the most common deep learning models used to detect toxic comments?

In this section, we want to analyze more about the deep learning algorithm used by each research paper. To do so, we have created a table. Table IV provides the method used, the preprocessing and performance of the algorithms.

Table IV. *Machine Learning and Deep Learning Algorithm*

Method	Reference	Preprocessing	Performance (%)
CNN	[11]	GloVe	Accuracy: 97.27%
	[12]	FastText	F1-Score: 0.8675%
	[19]	SSWE	Precision: 93% Recall: 94% F1-Score: 93%
	[23]	Word2Vec	Accuracy: 91.2% Specificity: 91.7% False disc rate: 8.3% Recall: > 90% F1-Score: > 90% Precision: >90%
	[10]	N/A	Accuracy: 98.93%
	[17]	GloVe, Word2Vec	7 Epoch Accuracy: 98.04%
	[3]	Glove Embeddings (100d)	Accuracy: 95% F1 Score: 87%
LSTM	[9]	LabelEncoder, SpaCy	Accuracy: 94.94% Precision: 94.49% Recall: 92.79%
	[11]	GloVe	Accuracy: 97.01%
	[13]	Gradient Descent	Accuracy: 79.1%
	[19]	SSWE	Precision: 91% Recall: 93% F1-Score: 88%
	[7]	GloVe, Topic, Sent	ROC_AUC: 91% Accuracy: 82.5% Macro F1: 83%
	[6]	N/A	F1-Score: 81%
	[17]	Word2Vec	7 Epoch Accuracy: 98.77%
	[24]	N/A	Total Positive Rate: 67% F1: 73% Precision: 81% Recall: 66%
Bi-LSTM	[12]	FastText	F1-Score: 85.98%
	[19]	SSWE	Precision: 92% Recall: 95% F1-Score: 93%
	[4]	Topic, GloVe	F1-Score: 66.7%
GRU	[3]	Glove Embeddings (100d)	Accuracy: 96% F1 Score: 90%
		Fasttext Embeddings (300d)	Accuracy: 96% F1 Score: 90%
Bi-GRU	[12]	FastText	F1-Score: 86.43%
NN	[17]	TF-IDF	7 Epoch Accuracy: 98.91%
DNN	[8]	AutoML, LEAF	AUROC: 84.3%
LSTM-CNN	[11]	GloVe	Accuracy: 96.95%
	[22]	GloVe, TensorFlow	Average Micro F1: 88.76% Average Macro F1: 67.98% ROC_AUC: 71%
	[10]	CountVectorizer	Accuracy: 98.39%
Logistic Regression	[18]	Multi-Label Classification	Hamming-Loss: 2.43% Accuracy: 89.46% Log-loss: 2.14%
	[25]	Regex Tokenizer, PorterStemmer, NLTK, Bag of Words	Accuracy: 92.1% Precision: 95.9% Recall: 92.0% F1-Score: 93.9%
	[16]	OneVsRest classifier	Accuracy: 97.1%
	[10]	N/A	Accuracy: 91%
	[21]	FastText300+Udemy reviews word embeddings	Accuracy: 96%
	[21]	GloVe300+Udemy reviews word embeddings	Accuracy: 96%
Naive Bayes	[24]	N/A	Total Positive Rate: 48% F1: 64% Precision: 94% Recall: 48%
Multinomial Naive Bayes Classifier	[3]	TF-IDF	Accuracy: 94% F1 Score: 82%
SVM	[14]	Text deduplication	Accuracy: 87.24%
Multi-Task Multi-Embedding Deep Learning Architecture	[2]	TF-IDF, FastText, Glove, Our	F1-Score (negative labels clean comment): 99% F1-Score (positive labels illegal comment): 59%
Ensemble Model	[1]	Meta-classifier	Accuracy: 93.2%
	[20]	Glove, Word2Vec	F1-Score: 86.6%

(Continued) Table IV. *Machine Learning and Deep Learning Algorithm*

Method	Reference	Preprocessing	Performance (%)
BERT	[26]	N/A	Micro F1: 78.16% Macro F1: 63.72%
	[6]	N/A	F1-Score: 81%
	[5]	N/A	Accuracy: 98.6%
RoBERTa	[26]	N/A	Micro F1: 78.8822 Macro F1: 65.10%
XLM	[26]	N/A	Micro F1: 75.94% Macro F1: 51.22%
Six-headed Machine Learning Model	[15]	TF-IDF	Accuracy: 98.08%

In [19], it was shown that deep neural networks are much more effective in detecting cyberbullying among three datasets. Other research also stated that deep learning gave better results compared to machine learning models [13]. Therefore, we are focusing more on deep learning algorithms in the context of text processing. In deep learning, there are three main algorithms that are largely used, such as CNN, RNN, LSTM, GRU which will be explained in detail below.

Usually, a machine learning algorithm learns from a dataset it is given to do a specific task, it recognizes patterns from the data given and makes predictions for new data. Deep learning algorithms are a subset of machine learning algorithms that are more complex, these algorithms consist of layers of algorithm to make a logic structure to draw a conclusion of the data given. The amount of data needed is more but there is minimal human intervention.

An example of a machine learning algorithm is logistic regression. Logistic regression is a machine learning algorithm that is used to predict whether something is true or false, after the model is fed with the training data to form an S-shaped curved line called a logistic function, this line can then be used to predict new data samples.

Convolutional Neural Network (CNN) is a neural network that is usually used to detect and identify an image and has a unidirectional model structure. During the training process, the neural network is fed with training data and processed with random weights inside the hidden layers, after the process is done, we compare the prediction result with the actual result and calibrate the weights inside the hidden layers. What makes CNN different from other neural networks are the Convolutional Layers and Fully Connected Layer. Convolutional and pooling Layers are used to extract the features from the dataset, while the Fully-Connected layers are used to shape the data we received from the feature extraction into a vector.

Although CNN is mostly used in Computer Vision, recent studies have also used CNN in natural language processing which has unprecedented results [11], CNN creates a row of fixed-dimension vectors for each word, where it can be used to make n-grams that seem like a one-way node of various sizes being passed over the words.

In terms of accuracy, CNN itself can be categorized as an algorithm that is able to compete with other algorithms because of its high accuracy value where it can be seen in Table 4 that CNN itself can gain more than 85 percent. Furthermore, it can be improved by using word embeddings like GloVe and Word2Vec to around 90% of accuracy.

Long Short-Term Memory (LSTM) is a special kind of recurrent neural network that has internal memory capacity to store and remember extended sequences, which is beneficial for text classification because it can recall long-term memory compared to RNN itself [19]. LSTM creates each hidden node into a memory cell where it can store additional data. Another advantage is LSTMs have input and forget gates, which allow for greater control of the gradient flow and improved maintenance of long-range dependencies [24]. These gates are how LSTM solves the RNN's problems, where they suffer in vanishing gradients [11].

Nowadays, LSTM is the most frequently used algorithm in deep learning. As a result of their high performance and accuracy value. In [11], LSTM is able to gain approximately 97% of accuracy which beats the accuracy score of other machine algorithms alone. That's what makes LSTM very widely used, especially in the field of Natural Language Processing (NLP).

Gated Recurrent Unit (GRU) is a special kind of recurrent neural network that has a gating mechanism and internal memory capacity, the mechanism that decides what information is passed to the output and removes or forgets irrelevant information. The GRU is similar to LSTM because it also solves RNN's vanishing gradient problem.

Models that include GRU with word embedding are one of the most effective models and are able to compete with other models in accuracy, it is proven that GRU can get a 96% accuracy score using GloVe or FastTex .[3] that is the reason why GRU is a popular choice in the field of Natural Language Processing.

Hybrid Model is combining two or more separate methodologies in order to improve accuracy. When compared to machine learning approaches, CNN+LSTM has a higher F1-Score [22]. However, when compared to single models based on datasets, the accuracy may be lower. [11] shows that CNN+LSTM+GloVe is less accurate than CNN+Glove or LSTM+GloVe alone.

It indicates that, aside from the model itself, not every hybrid model can attain improved accuracy. Some studies necessitate a more thorough examination in order to choose the optimal classification approach.

IV. Conclusion

In conclusion, LSTM or Long Short-Term Memory is the most used and consistently yields better accuracy in classifying toxic comments compared to other models with the highest at 98.77% accuracy depending on the pre-processing. GRU or Gate Recurrent Unit is also commonly used and yields good accuracy. While for Logistic Regression, the highest accuracy is only at 96%. However, we found that using Neural Networks with TF-IDF pre-processing methods yields the best accuracy for this task. Many researchers also try to propose new models by

combining two or three different methods to get a more accurate classification. However, hybrid models sometimes have lesser accuracy than the original model.

For datasets, we found that the most frequently used datasets are from Kaggle. Specifically, Wikipedia's talk page edits datasets that have over 150000 samples with a total of 13 out of 26 research papers using this particular datasets. We can deduce that researchers are recommended to use Wikipedia's talk page edits dataset from Kaggle in order to be able to accurately deduce the difference in evaluation scores between each algorithm. Furthermore, future studies might include more hybrid models with different models which have not been covered in this paper and researchers could use LIME model interpretability techniques to further help in finding the most reliable models rather than only depending on the evaluation metrics score for choosing the best model.

REFERENCES

[1]. A. Mohammed and R. Kora, "An effective ensemble deep learning framework for text classification," *Journal of King Saud University - Computer and Information Sciences*, 2021.

[2]. A. Elnaggar, B. Waltl, I. Glaser, J. Landthaler, E. Scepankova, and F. Matthes. "Stop Illegal Comments: A Multi-Task Deep Learning Approach." *In Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference (AICCC '18). Association for Computing Machinery, New York, NY, USA, 41–47, 2018*

[3]. A.Mahajan, D. Shah, and G. Jafar, "Explainable AI approach towards toxic comment classification." In Emerging Technologies in Data Mining and Information Security (pp. 849-858), 2021.

[4]. D. Y. Kim, X. Li, S. Wang, Y. Zhuo, and R. K. Lee. "Topic enhanced word embedding for toxic content detection in Q&A sites." *In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19). Association for Computing Machinery, New York, NY, USA, 1064–1071, 2019*

[5]. Fan, W. Du, A. Dahou, A. A. Ewees, D. Yousri, M. A. Elaziz, A. H. Elsheikh, L. Abualigah, and M. A. A. Al-qaness, "Social media toxicity classification using Deep Learning: Real-World Application UK Brexit,"(MDPI), 2021.

[6]. H. Almerekhi, H. Kwak, J. Salminen, and B. J. Jansen. "Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions." *Proceedings of The Web Conference 2020. Association for Computing Machinery, New York, NY, USA, 3033–3040, 2020*

[7]. H. Almerekhi, Supervised by B. J. Jansen, and co-supervised by H. Kwak. "Investigating Toxicity Across Multiple Reddit Communities, Users, and Moderators." *Companion Proceedings of the Web Conference 2020. Association for Computing Machinery, New York, NY, USA, 294–298, 2020*

[8]. J. Liang, E. Meyerson, B. Hodjat, D. Fink, K. Mutch, and R. Miikkulainen. "Evolutionary neural AutoML for deep learning." *In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '19). Association for Computing Machinery, New York, NY, USA, 401–409, 2019*

[9]. K. Dubey, R. Nair, M. U. Khan, and P. S. Shaikh, "Toxic comment detection using LSTM," *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, 2020.

[10]. M. A. Saif, A. N. Medvedev, M. A. Medvedev, and T. Atanasova, "Classification of online toxic comments using the logistic regression and neural networks models," ,2018.

[11]. M. Anand and R. Eswari, "Classification of abusive comments in social media using Deep Learning," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019

[12]. M. Ibrahim, M. Torki, and N. El-Makky, "Imbalanced toxic comments classification using data augmentation and Deep Learning," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018.

[13]. M. Mahat, "Detecting cyberbullying across multiple social media platforms using Deep Learning," *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021.

[14]. M. Zhang, "E-Commerce Comment Sentiment Classification Based on Deep Learning." *IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 2020, pp. 184-187, 2020*

[15]. N. Chakrabarty1, "A Machine Learning Approach to Comment Toxicity Classification." , 2019

[16]. P. Ravi, H. N. Batta, S. Greeshma, S.Yaseen, "Toxic Comment Classification" , 2019

[17]. R.Sharma , & M.Patel, "Toxic Comment Classification Using Neural Networks and Machine Learning" - IARJSET. Academia, 2018

[18]. Rahul, H. Kajla, J. Hooda, and G. Saini, "Classification of online toxic comments using machine learning algorithms," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020.

[19]. S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," *Lecture Notes in Computer Science*, pp. 141–153, 2018.

[20]. S. Alshamrani, A. Abusnaina, M. Abuhamad, D. Nyang, and D. Mohaisen. "Hate, Obscenity, and Insults: Measuring the Exposure of Children to Inappropriate Comments in YouTube." *In Companion Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 508–515, 2021*

[21]. S. Carta, A. Corriga, R. Mulas, D. Recupero, and R. Saia, "A Supervised Multi-class Multi-label Word Embeddings Approach for Toxic Comment Classification," *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2019

[22]. S. Gilda, L. Giovanini, M. Silva, and D. Oliveira, "Predicting different types of subtle toxicity in unhealthy online conversations," *Procedia Computer Science*, vol. 198, pp. 360–366, 2022.

[23]. S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, and V. P. Plagianakos. "Convolutional Neural Networks for Toxic Comment Classification." *In Proceedings of the 10th Hellenic Conference on Artificial Intelligence. Association for Computing Machinery, New York, NY, USA, Article 35, 1–6, 2018*

[24]. S. Zaheri, J. Leath, and D. Stroud, "Toxic Comment Classification" , SMU Data Science Review: Vol. 3: No. 1, Article 13, 2020

[25]. V. Jain, V. Kumar, V. Pal, and D. K. Vishwakarma, "Detection of cyberbullying on social media using machine learning," *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021.

[26]. Z. Zhao, Z. Zhang, and F. Hopfgartner. "A Comparative Study of Using Pre-trained Language Models for Toxic Comment Classification." *In Companion Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 500–507, 2021*