Toxic Comment Identification and Classification using BERT and SVM

Ivander Gladwin
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
ivander.gladwin@binus.ac.id

Evan Vitto Renjiro
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
evan.renjiro@binus.ac.id

Bryan Valerian
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
bryan.valerian@binus.ac.id

Ivan Sebastian Edbert
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
ivan.edbert@binus.ac.id

Derwin Suhartono
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
dsuhartono@binus.edu

2022 8th International Conference on Science and Technology (ICST) | 978-1-6654-5622-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICST56971.2022.10136295

***Abstract*— Bullying cases like toxic comments on many social media platforms cause a negative impact that occurs in every age circles. From those cases, we would like to make a system that can identify and classify toxic words from a comment before it is sent and seen by others. By utilizing a Machine Learning application, hopefully, the produced system can be useful in reducing bullying cases that are many in social media. Lot of experiments have been done to find the settlement for this problem, but various algorithms and models are used. In this research, we will be doing a comparison of two models, the BERT (Bidirectional Encoder Representations from Transformers) model which is usually used to solve NLP (Natural Language Processing) tasks, and SVM (Support Vector Machine) model which is great at classifying. Both models will be compared to find out which model is better in identifying and classifying toxic comments. The result that is gotten shows that BERT model is said to be superior compared to SVM model, with an accuracy of 98.3% including other metric evaluation scores that show a significant result compared to the result achieved by SVM model.**

Keywords—Machine Learning, Toxic Comments, Support Vector Machine, Natural Language Processing, Transformer Model

I. INTRODUCTION

In the present era, almost everyone can access social media. Social media is an online media where users can give opinions or comment on other users. However, this matter is always misapplied, and it becomes a problem where users love to give a toxic comment to other users [1]. A toxic comment is a comment that is intended to vilify someone by using various means, including violence when chatting. Comment moderation actually can be done manually, but to do this thing may cost a big sum of money, be less effective, and sometimes inappropriate. If toxic comments can be identified automatically, we can have a safer discussion on any kind of social network, news portal, or even online forum [2].

Recently, there are many cases where hatred and negativity have evolved and are mostly seen on online platforms, especially social media. Based on research that was done by Microsoft on Mei - April 2020 with the "Digital Civility Index" as a benchmark, Indonesia's netizens are listed at the 29th or bottom three as "disrespectful netizens". This research is done in 32 countries with a total of 16.000 respondents, 503 of them from Indonesia. In Indonesia itself, the most frequent cyberbullying that happens is hoax spread and fraud (47%), hate speech (27%), and discrimination (13%).

In 2020, a survey was done by U-Report Indonesia that involves 2.777 Indonesian respondents with a 97% rating for the response. The result from the respondents shows that 45% of people have encountered digital violence. Furthermore, digital violence often happens on social media with a percentage of 71%, 19% in chatting applications, 5% in online games, 1% on YouTube, and 4% from the others that are not mentioned. From 97% of total respondents, 34% respondents being a victim unwilling to get service or help and the other 36% don't know any information regarding cyberbullying service centers. Other data shows that 39% of netizens feel that the government is the one that should be responsible for cyberbullying cases, followed by 11% for schools, 14% for the internet service providers, and 36% for young kids [3].

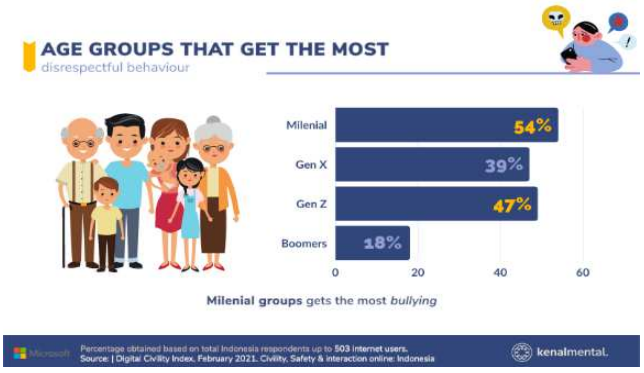

Figure 1 Data Age Group most frequently exposed to cyberbullying

Reported from the research data as shown in Figure 1, it is known that the age group that often gets cyberbullying is the millennial group (1980-1995) with a percentage of 54%, followed by the Z generation (1997-2012) with a percentage of 47%, then X generation (1965-1980) with a percentage of

978-1-6654-5622-7/22/$31.00 ©2022 IEEE

39%, and lastly baby boomers (1946-1964) with a percentage of 18%. According to a report from Polda Metro Jaya, there are at least 25 cyberbullying cases that are reported daily. This number will keep increasing due to the number of internet users and the less effective way of handling cyberbullying cases [4].

With many papers that classify and identify toxic comment using different methods, this paper will explain how we classify and identify toxic comment that is inputted by users into some categories such as toxic, obscene, insult, severe-toxic, identity-hate, and threat comment using pre-trained language BERT model. We also want to use the SVM model for comparing both models and determining the performance of each model.

II. LITERATURE REVIEW

Quite many studies have been conducted to detect and classify toxic comments during the last 5 years with various models and algorithms. An experiment using NLP, LTSM (Long Short-Term Memory), and CNN (Convolutional Neural Networks) to identify toxicity in the text shows a result where the best model for word-level binary classification is CNN [5]. On the other hand, LTSM performed highest in multi-label and binary classification, achieving the highest accuracy, precision, and F1-score among all models that were tested. Another research is conducted, but using a different model, which is Logistic Regression (LR) as the model to train the data and confusion matrix to summarize the performance of the classification algorithm [6]. As the result, they want to produce a system that can automatically classify a comment with more than 95% accuracy for each label category. Next, there are different models, where word embedding techniques and Recurrent Network Neural (RNN) are used [7]. In her result, it was shown that there are more models used such as Gated Recurrent Units (GRU), pre-trained word embedding vectors, penalizing loss, and undersampling that will be used to compare with the baseline model, word2vec embedding with biLTSM (Bidirectional LSTM). The comparison shows that the GRU layer proves to be more efficient at training but performs slightly worse than the baseline model.

More research show how to classify with various models, such as Deep Neural Network (DNN) architecture to solve the problem of overlapping toxic sentiment classification [8]. The results show that based on the results of performance metrics, the Bidirectional GRU is the best model to solve this case, with a note that doesn't spend too much time in data preprocessing. Different from the others, there is research aimed to classify unstructured text into toxic and non-toxic categories by using Naïve Bayes and LSTM/RNN algorithm [9]. The result shows that by using the Naïve Bayes method, 745 toxic comments out of 1,543 toxic comments were identified. On the other hand, LSTM/RNN algorithm were able to identify 1,026 toxic comments out of 1,543 toxic comments. It was concluded that LSTM/RNN algorithm is better at classifying toxic comments than Naïve Bayes Method, with a difference of almost 20% for the true positive rate. Toxic comment classification using NLP with several methods such as LSTM, CNN, and Naïve Bayes Support Vector Machine (NB-SVM) and also use Fasttext separately is also conducted to see which model fits and works better than the others [10]. After analyzing various approaches, CNN models is concluded as the best model where it works slightly better than LSTM and NB-SVM with an accuracy of 98.13%. Transformers model as one of the models used to develop a classification of toxic comments including more models to be compared shows a result that transformer model outperformed other models with a fairly visible difference in results [11]

Modification or combination of one or more models are also executed, for example SVM with a decision tree as a model and the other model is Neural Networks [12]. By applying the TF-IDF (Term Frequency-Inverse Document Frequency) feature extraction method and neural network techniques, it outperforms the other 2 models with 97.07% accuracy. There is also GRU by using SGDM optimizer [13]. From the experiments, it is found that the GRU model using SGDM optimizer achieved the best performance F1-score of 0.978. result also shows that combined Neural Network achieved more than 90% effectiveness and the best combination was GRU layer as the first layer and Bi-LSTM as the second layer.

To enhance our knowledge regarding SVM that are generally used to classify toxic comments, we search for more literature regarding the SVM model. One of the experiments use several models such as BR Method with MNB Classifier, and BR Method with SVM Classifier [14]. When measuring hamming loss as a benchmark for identifying the optimal algorithm for classifying toxic comments, the Binary Relevance method with MultinomialNB is the best algorithm for classifying the toxic comments with a hamming loss score of 3.6 compared to the hamming loss of SVM of 4.36. SVM is also used in other experiments, including linear, RBF, Sigmoid, and Chi SQUARE Feature Selection to classify the toxic comments [15]. From the result that they already experiment with, using SVM as a model with linear kernel and without using the Chi Square feature has the highest F1 score with 76.57% outperforming other models.

Related to the algorithm that we are using, which is BERT, NB SVM method is used as the basline model to be compared with BERT, and other algorithms in the research [16]. Research shows that BERT has the best results even with just one epoch and by adding a very good Weighted Loss correct data imbalances. There is also comparison among various BERT models, from BERT, Multilingual BERT, RoBERTa, to DistilBERT [17]. Result found that BERT can classify and predict toxic comments with a high degree of accuracy which was 0.98603 score compared to the other models. BERT is also used to be compared, with its friend RoBERTa and another different model which is XLM [18]. The research found that BERT and RoBERTa perform better than using XLM when they do the classification tasks. Lastly, a different BERT is used, which is German BERT being compared with Transformers Multi-Layer Perceptron [19]. With many fine-tuned models that are already being trained with over 1.5 million data, it is found that German BERT still has stronger predictions. From this, we can know that the BERT model with outperform other models in term of performance to classify toxic comments.

III. METHODOLOGY

In this toxic comment classification, we are using a dataset provided by Competition Toxic Comment Classification Challenge on Kaggle that contains toxic comments from the year 2004 to 2015, taken from many forums such as Civil Comments and Wikipedia Talk Pages. CC-SA 3.0 Wikipedia has managed the existing data text and is managed under CC0. The toxic comment itself will be divided into different toxic categories starting from toxic, obscene, insult, severe-toxic, identity-hate, and threat comment [20]. The data from each toxic comment category can be seen in Figure 2.

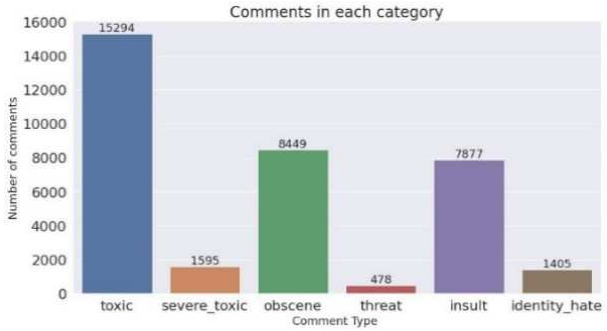

Figure 2: Comments in Each Category

Toxic comment data from Kaggle have been divided into 6 categories which we can see in the example in Table 1. All data can be divided into more than 1 category or even not categorized into any category at all.

A. Proposed Model

In this paper, we will be using 2 methods which are BERT and SVM by using Google Collab as a tool to help in running the code. The metrics of measuring that will be used to calculate accuracy from the classification are accuracy, AUC, and Logloss. The purpose of using these 2 models are to know each performance to conclude which model is more accurate in classifying toxic comment. Tuning won't be done in this paper as we are trying to compare the performance of both models from the origin, without any changing.

NLP itself is a branch of AI that deals with interactions between computers and humans using natural language [21]. NLP is used to measure sentiment and determine which parts of human language are important. Most likely, we have worked alongside NLP in various forms and interactions such as GPS systems that are operated with voice, speech recognition that will be changed into text, and chatbots in customer service features. Not only that, NLP also has a role in business solutions such as increasing employee efficiency in working and helping streamline business operations [22].

BERT is a deep learning algorithm that has been around since 2019 and has been designed to process Natural Language Processing (NLP) [23]. Corresponding with the name, BERT only encodes and produces a language model [24]. It caused quite a stir in the Machine Learning community by presenting cutting-edge results in various NLP activities like Natural Language Inference (MNLI), Answering Questions (SQuAD v1.1), etc [25], [26]. The architecture of the BERT model is based on Transformers and uses a multilayer bidirectional transformer encoder to do language representation. The BERT type used is the Pretrained

We used $BERT_{base}$ model, which operates 12 layers of transformer blocks with a hidden size of 768, a total of 12 self-attention heads, and has a trainable parameter of around 110M [27]. BERT achieve advanced performance for eleven NLP tasks, including Question Answerer and Sentence (and sentence pair) classification tasks, simply by refining the last layer. BERT has the noteworthy achievement of learning more powerful bidirectional representations than most previous approaches. BERT architecture can be seen in Figure 3.

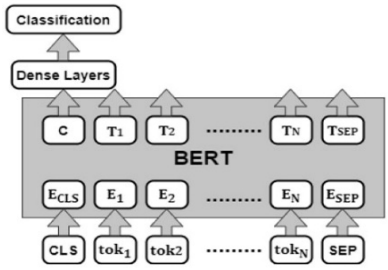

Figure 3: Architecture of BERT Model [34]

The BERT architecture is primarily a two-way, multi-layer Transformer encoder with a two-way self-attention mechanism. BERT uses two special tokens [CLS] and [SEP], to understand each input sequence properly [28]. [CLS] is the token that contains the special classification and the last hidden state of BERT which representation of the whole sequence input for the classification tasks. The [SEP] is to separate input segments and must be inserted at the of a single input [29].

The SVM model is a supervised machine learning model that can solve two group classification tasks using a classification algorithm. By providing the model with

Table 1 Random Sample form Datasets

id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0007e25b2121310b	Bye!\nDon't look, come or think of comming back! …	1	0	0	0	0	0
b2d7c107fdcb95fa	this guy is a dirty 213.152.254.36	1	0	0	0	0	1
d6f28744b607fcfa	You people are fucking morons\nStop hand-wrin …	1	0	1	0	1	0
001e89eb3f0b0915	Are you threatening me for disputing neutrality? …	0	0	0	0	0	0
001810bf8c45bf5f	You are gay or antisemmitian?\n Arc…	1	0	1	0	1	1

training data in each category, the SVM model can classify text well with better performance and higher accuracy compared to other classification algorithms [30].

There is a special feature in SVM that can reduce the empirical misclassification while maximizing the geometric margins where "maximum margin classifier" is named for this feature. In this method, the input vector will be mapped into a higher dimensional space vector through the predefined hyperplanes [31]. Hyperplane itself is a line that has been determined by the model to determine the limits of data classification. The dots between the lines already indicate if they are in different classes. In Figure 4, we can see a hyperplane that separates 2 classes [30], [32]. The points that are closest to the hyperplane are support vectors and they are used by the SVM to maximize the margin of the classifier. We will use SVC from sklearn svm library.

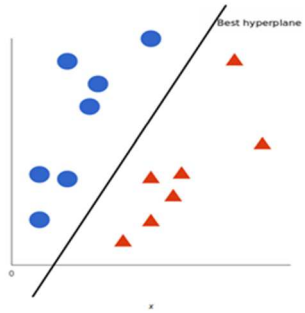

Figure 4: Best Hyperplane for Classify

B. *Proposed Architectures*

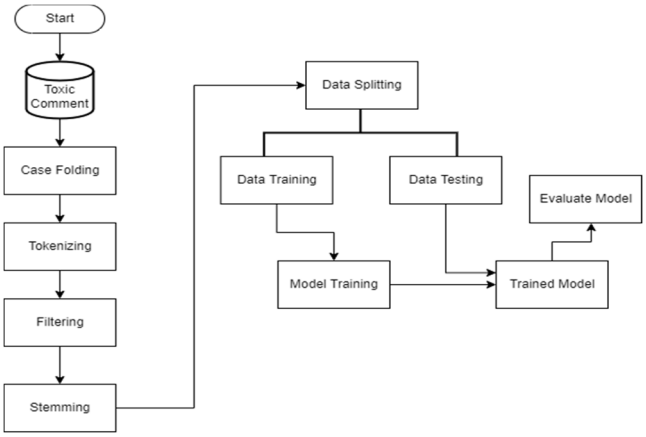

Figure 5 Workflow

Based on Figure 5, this research will have some stages in classifying the toxic comment process.

1. Case Folding
In this stage, all datasets will change each of the uppercase letters into a lowercase letter which will be helpful in toxic comment classification so that it doesn't add a new case where there is the same word but is defined as different just because of the difference on the uppercase letter.

2. Tokenizing
In this stage, a dataset that has already been case folding will be erased all its URL, hashtags, and punctuations inside the sentence from the dataset, and a separation of each sentence into a single word.

3. Filtering
In this stage, the deletion of the word that is defined as meaningless will be done using a stop word list.

4. Stemming
This stage is to reduce the different indexes from one data so that words that have suffixes or prefixes will return to their basic form.

5. Data Splitting
In this stage, data that is already available will be divided into 2 types that will be used as data training or even data testing to model that will be tested. Existing data will be divided into 70% for training and 30% for testing.

IV. Result and Discussion

In our experiment, we will be using some evaluation metrics to measure how well a model is used to classify toxic comments based on data that has been provided. Evaluation metrics that will be used are Accuracy, AUC, and Logloss from 2 models that are BERT and SVM. ROC is a measurement performance tool for classification problems to decide the threshold of a model. On the other hand, the AUC is an area below the ROC curve that can measure the performance of the classifier without the need for a certain threshold [33]. On the BERT model, it can be seen from Figures 6 and 7 below that display numeric graph of logloss score and AUC score from executed epoch.

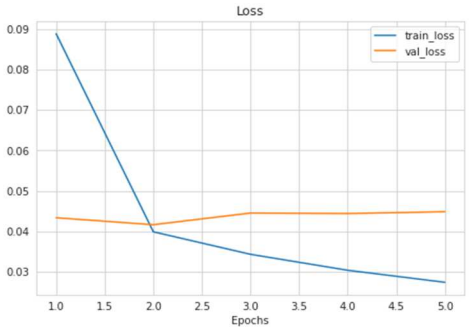

Figure 6: Log Loss for BERT in 5 epochs

Figure 6 represents logloss from 5 epochs using the BERT model, it can be seen that logloss training has a slope up to the fifth epoch with a result of 0.0416. From the statement before, it can be shown that the model used as training for classification can comprehend provided data very well. Also, from the picture itself can be seen that the logloss validation is quite stable from the third epoch to the fifth where the logloss shows stability in the third epoch.

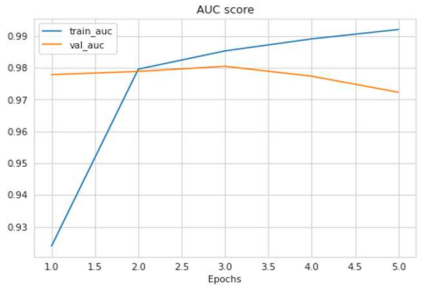

Figure 7: AUC Score for BERT in 5 epochs

In Figure 7 which represents the AUC score of 5 epochs done by the BERT model, it is known that the score keeps improving, however when it touches the third epoch up to the fifth, the AUC score that is achieved drops even though not significant. This can happen due to the AUC train score being very high, resulting in an AUC validation score that experiences a slope. The final score that is obtained from the AUC score shows a number of 0.9753 where occurs a bit degression compared to the AUC score on the first epoch which is 0.9760

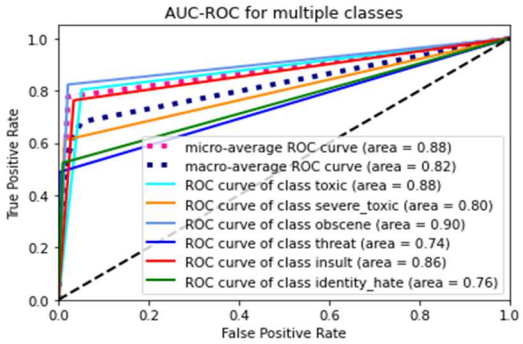

Figure 8: AU Figure 8: AUC-ROC for SVM model

On the other hand, the result of the classification from the SVM model can be seen in the graph that is displayed in figure 8. SVM itself classifies toxic comment by each category that are available which make the different result that is obtained in accordance with the category that is being classified. From figure 8, it can be seen that the ROC curve with the highest score was achieved when SVM classified the toxic comment in the Obscene category, while the lowest ROC score was achieved when classified in the Threat category. Here, we take the biggest score as a pole to decide how accurate the SVM model can classify provided toxic comments. In table 2, we can see the result of the classification calculation that is done by SVM for each toxic comment category.

Table 2 Result of AUC-ROC in SVM

Toxic Comment Categories	AUC-ROC
Toxic	88%
Severe_toxic	80%
Obscene	90%
Threat	74%
Insult	86%
Identity_hate	76%

Table 2 exhibits 6 classes that are classified by SVM, where the highest score is achieved by the Obscene class with a value of 90% and followed by the Toxic class. From this data, it can be shown that those 2 classes have a higher true positive rate for the sample probability than the lower false positive sample that is chosen randomly. The lowest score from this classification is gotten when SVM classified threat class and also identity_hate class. The nearer a score of 1 or 100% to the AUC-ROC score, the better the model is in deciding True Positive Rate to calculate other evaluation metrics. Based on table 2, AUC from the BERT model is far superior to the SVM model, because it shows that every point in the BERT model has a higher True Positive and or lower False Positive than SVM. The result gotten from classifying toxic comments using both models can be seen in Table 3 below.

Table 3 Result of the testing model

	BERT Model	SVM Model
Accuracy	**98.32%**	87.427%
AUC	**0.9753**	0.8233
Log-loss	**0.0416**	1.034

In Table 3, result for the BERT model is achieved by using 5 epochs while SVM model's result is gotten from the average result of classifying each toxic comments category. We can see that the BERT model is far superior to SVM model, with a different of more than 0.15 on the AUC, even showing a high accuracy almost reaching 100% and log-loss of less than 0.1. On the other side, SVM model shows a quite big value of log loss, compared to BERT, where the log-loss value of SVM model is reaching 1.

V. CONCLUSION

Research and experiments that we have done, which is comparing models for classifying and identifying toxic comments where the dataset itself comes from Kaggle dataset with the proportion of 70:30 data and is classified using BERT model and SVM. From the experiment, the result shows that using the BERT model gets a better result than SVM with an accuracy of 98.32%, AUC with a score of 0.9753, and logloss with a number of 0.0416.

For future work, we will use the dataset provided by ourselves in several language and create several new categories in the classification of toxic comments to get better results. We will also use another model and algorithms for comparison with the BERT model so that several choices of models that work well will be obtained that can be used to identify toxic comments on social media.

REFERENCES

[1] A. S. Cahyono, "Pengaruh Media Sosisal Terhadap Perubahan Sosial Masyarakat di Indonesia," pp. 140–157, Mar. 2017.

[2] D. Andročec, "Machine learning methods for toxic comment classification: a systematic review," *Acta Universitatis Sapientiae, Informatica*, vol. 12, no. 2, pp. 205–216, Dec. 2020, doi: 10.2478/ausi-2020-0012.

[3] "Cyberbullying: Racun Social Media di Indonesia," *Profesi UMN*, Nov. 29, 2021. https://profesi-unm.com/2021/11/29/cyberbullying-racun-social-media-di-indonesia/ (accessed Apr. 03, 2022).

[4] I. Admin, "Laporkan Polisi Bila Aksi Cyberbullying Menimpa Anak Kita," *Nanegeriku*, Mar. 16, 2021. https://inanegeriku.com/2021/03/16/laporkan-polisi-bila-aksi-cyberbullying-menimpa-anak-kita/ (accessed Apr. 04, 2022).

[5] K. Khieu and N. Narwal, "Detecting and Classifying Toxic Comments," 2017.

[6] P. A. Ozoh, A. A. Adigun, and M. O. Olayiwola, "Identification and Classification of Toxic Comments on Social Media using Machine Learning Techniques," 2019. [Online]. Available: www.rsisinternational.org

[7] S. Li, "Application of Recurrent Neural Networks In Toxic Comment Classification," 2018.

[8] H. H. Saeed, K. Shahzad, and F. Kamiran, "Overlapping toxic sentiment classification using deep neural architectures," in *IEEE International Conference on Data Mining Workshops, ICDMW*, Feb. 2019, vol. 2018-November, pp. 1361–1366. doi: 10.1109/ICDMW.2018.00193.

[9] S. Zaheri, J. Leath, and D. Stroud, "Toxic Comment Classification," 2020.

[10] A. A. Sagar and J. S, Kiran, "Toxic Comment Classification using Natural Language Processing," *International Research Journal of Engineering and Technology*, 2020, [Online]. Available: www.irjet.net

[11] G. Akash, H. Kumar, and D. Bharathi, "Toxic Comment Classification using Transformers," 2021.

[12] R. Patel and H. Gaudani, "Toxic Comments Classification using Neural Network," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 7S, pp. 12–15, May 2020, doi: 10.35940/ijitee.g1005.0597s20.

[13] K. Machová, M. Mach, and M. Vasilko, "Recognition of Toxicity of Reviews in Online Discussions," 2022. [Online]. Available: http://www.kaggle.com.

[14] P. Vidyullatha, S. N. Padhy, J. G. Priya, K. Srija, and S. Koppisetti, "Identification and Classification of Toxic Comment Using Machine Learning Methods," 2021.

[15] N. S. Azzahra, D. T. Murdiansyah, and K. M. Lhaksmana, "Toxic Comment Classification on Social Media Using Support Vector Machine and Chi Square Feature Selection," *Intl. Journal on ICT*, vol. 7, no. 1, pp. 64–76, 2021, doi: 10.34818/ijoict.v7il.552.

[16] H. Li, W. Mao, and H. Liu, "Toxic Comment Detection and Classification," 2019.

[17] H. Fan *et al.*, "Social media toxicity classification using deep learning: Real-world application uk brexit," *Electronics (Switzerland)*, vol. 10, no. 11, Jun. 2021, doi: 10.3390/electronics10111332.

[18] Z. Zhao, Z. Zhang, and F. Hopfgartner, "A Comparative Study of Using Pre-Trained Language Models for Toxic Comment Classification," in *The Web Conference 2021 - Companion of the World Wide Web Conference, WWW 2021*, Apr. 2021, pp. 500–507. doi: 10.1145/3442442.3452313.

[19] M. Schütz *et al.*, "DeTox at GermEval 2021: Toxic Comment Classification," 2021. [Online]. Available: https://pypi.org/project/emosent-py/

[20] J. Peng *et al.,* "Toxic Comment Classification Challenge," *Kaggle*, Dec. 19, 2017. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (accessed Apr. 03, 2022).

[21] M. Koroteev, *BERT: A Review of Applications in Natural Language Processing and Understanding*. 2021.

[22] "Natural Language Processing (NLP)," *IBM Cloud Education*, Jul. 02, 2020. https://www.ibm.com/cloud/learn/natural-language-processing (accessed Apr. 20, 2022).

[23] L. Ben, "BERT Language Model," *TechTarget*. Apr. 2020. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model

[24] F. A. Pratama and A. Romadhony, "Identifikasi Komentar Toksik Dengan BERT," 2020.

[25] H. Rani, "BERT Explained: State of the art language model for NLP," *Towards Data Science*, Nov. 11, 2018. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (accessed Apr. 19, 2022).

[26] Z. Zhao, Z. Zhang, and F. Hopfgartner, "A Comparative Study of Using Pre-Trained Language Models for Toxic Comment Classification," in *The Web Conference 2021 - Companion of the World Wide Web Conference, WWW 2021*, Apr. 2021, pp. 500–507. doi: 10.1145/3442442.3452313.

[27] K. Raman, "All You Need to Know About BERT," *Analytics Vidhya*, May 27, 2021. https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-bert/ (accessed May 09, 2022).

[28] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works," 2020, doi: 10.1162/tacl.

[29] H. Li, W. Mao, and H. Liu, "Toxic Comment Detection and Classification," 2019.

[30] S. Styawati and K. Mustofa, "A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 3, p. 219, Jul. 2019, doi: 10.22146/ijccs.41302.

[31] C. Dias and M. Jangid, "Vulgarity Classification in Comments Using SVM and LSTM," in *Smart Systems and IoT: Innovations in Computing*, 2020, pp. 543–553.

[32] B. Stecanella, "Support Vector Machines (SVM) Algorithm Explained," *MonkeyLearn*, Jun. 22, 2017.

[33] B. V. Aken, J. Risch, R. Krestel, and A. Löser, "Challenges for Toxic Comment Classification: An In-Depth Error Analysis," Sep. 2018, [Online]. Available: http://arxiv.org/abs/1809.07572

[34] S. Hanane and E. F. Nour-eddine, "Fine-Tuned BERT Model for Large Scale and Cognitive Classification of MOOCs," *RiME Team, MASi Laboratory, E3S Research Center, Mohammadia School of Engineers (EMi), Mohammed V University, Rabat, Morocco*, May 2022.