OPEN ACCESS
ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol.7, Issue.1, June 2021. pp. 64-76
doi:10.34818/ijoict.v7il.552

Toxic Comment Classification on Social Media Using Support Vector Machine and Chi Square Feature Selection

Nadhia Salsabila Azzahra [1], Danang Triantoro Murdiansyah [2*], Kemas M. Lhaksmana[3]

[123]*School of Computing, Telkom University*
Bandung 40275, Indonesia

[*] danangtri@telkomuniversity.ac.id

Abstract

The use of social media in society continues to increase over time. The ease of access and familiarity of social media makes it easier for an irresponsible user to do unethical things such as spreading hatred, defamation, radicalism, pornography, etc. Although there are regulations that govern all the activities on social media, it is still not working effectively due to the impossibility of classifying the comments manually. Therefore, we conducted this study to classify comment into their toxicity categories using machine learning methods for convenience purposes on social media usage. The method that we used in this study is SVM with TF-IDF as the feature extraction and Chi Square as the feature selection. We also performed several exploration scenarios, including implementing SVM kernels and preprocessing stages to find out the best performance of the model. The best performance obtained using the SVM model with a linear kernel, without implementing Chi Square, and using stemming and stopwords removal with the F 1 − Score equal to 76.57%.

Keywords: text classification, toxic comment, social media, support vector machine.

I. INTRODUCTION

In this digital era, information and communication technology is developing rapidly, including social media. Social media is one of the means that a person uses to interact with each other by giving, sharing, and exchanging information or ideas in a virtual network [1]. Based on a survey conducted by Hootsuite, in 2020, internet usage in Indonesia increased by 17% compared to the previous year, reaching 175.4 million people of which 160.0 million were active users on social media such as YouTube at 88%, WhatsApp by 84%, Facebook at 82%, Instagram at 79%, and Twitter at 56% [2].

By the existence of social media, everyone has the right to express their opinions freely with minimum restrictions. It could lead to social media misused by irresponsible peoples, either a person or a group of people, to spread hatred, racist comments, radicalism or extreme ideology, pornography, defamation, and so on, even though there is now a legal that limits social media usage. This phenomenon leads some of the researchers to distinguish some different types of toxicity in a comment to avoid undesirable things.

There were several studies on the classification of toxic comments that have conducted before. Previous research classified tweet data using the Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction and using the Support Vector Machine (SVM) method and compared the model with the

Received on 10 May 2021, Revised at 7 Jun 2021, Accepted at 11 Jun 2021, Published online on 2 Jul 2021

Convolutional Neural Network model [3]. The best $F1-Score$ obtained using the TF-IDF and SVM models equal to 74.88% due to SVM can classify data with two labels with only a few appearances. The preprocessing stages that yield the best result are without stopwords removal, stemming, and translation. Furthermore, detecting hateful comments on YouTube and Facebook social media using TF-IDF as feature extraction and the linear-SVM method obtained 79% of $F1-Score$ [4]. Classification of text in Arabic using Chi Square, Information Gain, and Mutual Information compared feature selection using two different methods, namely SVM and Decision Tree. The best $F1-Score$ value is obtain using Chi Square and SVM method, which is 79.59%. Using Chi Square, the calculation process in classification becomes faster [5].

In this study, we used the SVM method. Since, this method is proven to deliver good performance in the previous research such as in [3], [4], and [5]. Besides, we also use three different kernels for SVM, including Linear, RBF, and Sigmoid to find out which kernel obtain the best performance. Furthermore, we use TF-IDF as feature extraction and Chi Square as feature selection. The test scenario in this studyis with and without using Chi Square for each SVM kernel referred to [5], while the preprocessing stageis to compare the use of stemming and stopwords and without using stemming and stopwords referred to [3]. The evaluation is using the $F1-Score$.

The rest of the paper is organized as follows. Section II is the literature review. Followed by sectionIII describes the methodology we used in this research. Furthermore, Section IV presents experimental results on the dataset using several scenarios. Lastly, Section V describes our conclusions and future works based on our experiment results.

II. LITERATURE REVIEW

Classification of toxic comment is a well-known task, especially toxic comments in English. Unfortunately, this kind of research in Bahasa Indonesia is still rarely done due to the lack of data. Classify toxic comments using NBSVM, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). The NB-SVM model uses TF-IDF as its extraction feature. The best accuracy is obtained by using the NB-SVM model, with accuracy is equal to 98.13% [6]. Another research performed multilabel classification using several Binary Relevance classification methods such as Naïve Bayes, SVM, and KNN and using several feature selections such as Mutual Information, Odd ratio, and Chi Square. The best $F1-Score$ is obtained using the SVM method with Chi Square as its feature extraction is equal to 80.03% [7]. Moreover, classify toxic comment in Urdu and implement several experiments using several machine learning and deep learning models, namely Naïve Bayes, SVM, Random Forest, Logistic Regression, KNN, LSTM, and CNN. The best performance is using the Random Forest model with 0.966 of $F1-Score$ [8].

There are relatively a few kinds of research on toxic comment classification in Bahasa Indonesia. In 2019, the multilabel hate speech and abusive language classification used NB, SVM, and Random Forest Decision Tree (RFDT) as the models. Furthermore, implemented several data transformations such as CC, LP, and BR are implemented. The best accuracy obtained was by using RFDT and LP to identify abusive language and hate speech without identifying the target, categories, and level of hate speech with the accuracy equal to 77.36% and 66.12% of accuracy by identifying the target, categories, and level of hate speech [9]. The research conducted by [3] is to classify tweets that contain elements of hate speech using several methods, including SVM, CNN, and DistilBERT. The best is using SVM and without stemming, and stopwords removal, which is 74.88%. Meanwhile, the lowest $F1-Score$ is using CNN.

A. Multilabel Classification

Multilabel classification is one of a Natural Language Processing (NLP) task to categorize data into certain classes [10]. This classification has been used in several cases genbase classification [11], text [7], music genre [12], etc. The methods that we could implement for multilabel classification also vary, starting from using machine learning methods, such as SVM, Naïve Bayes, and KNN to use deep learning methods, such as LSTM and GRU and models hybrid.

B. TF-IDF

Term Frequency - Inverse Document Frequency (TF-IDF) is one of the weighting methods in feature extraction or feature selection which is often used in information retrieval because of its good performance [13]. This method assessed the importance of an existing word. If a word appears more often on a document, the value of its contribution will be even higher. However, if the word only appears in a few documents, then the resulting contribution's value was lower. TF-IDF combines two different methods, namely, TF, which functions to count the occurrence of the same word in a document, and IDF, which works to count the number of similar documents containing specific words [14]. The formula of TF-IDF shown in Equation 1.

$$TF - IDF(t, d) = tf_{t,d} \times log\left(\frac{N}{df_t}\right) \quad (1)$$

Where $tf_{t,d}$ is the frequency of the word t occurs in dataset d, N is the number of documents in the dataset, and lastly df_t is the number of document that contain word t. In TF-IDF the weighting of words that appear more frequently in a document will have a greatercontribution value, and vice versa [14].

C. Chi Square

This method obtains dependent features by calculating the distribution to measure the dependency value between features and classes. The calculation for the Chi Square function is shown in Equation 2.

$$x^2(t, c) = \frac{N(AD - CB)^{\angle}}{(A + B)(B + D)(A + B)(C + D)} \quad (2)$$

Where *t* is a word from *c* class and *N* is the total number of documents contained in the trainingdata. Furthermore, *A* the number of documents in class *c* which contains the word *t*, *B* the number of documents that are not in the class *c* and contains the word *t*, *C* the number of documents contained in class *c* but does not contain the word *t* and *D* is the number of documents that are not class *c* and doesnot contain the word *t* [15].

D. SVM

Support Vector Machine (SVM) is a classification technique that can classify data non-linearly. SVM works in defining a hyperlane with the functions to separate a data set into two different classes. The most optimum hyperlane value is when the hyperlane is in the middle, between two classes separated by the hyperlane [16]. In SVM, the general hyperlane function is shown in Equation 3.

$$H: w^T(x) + b = 0 \quad (3)$$

Where w is the vector of attribute w, x is the vector of the attribute x and b is the intercept and bias term of the hyperlane. Furthermore, the value of x and b can be positive, zero, or negative.

Practically, SVM can only classify data linearly. Therefore, to overcome this problem, we can use the features provided by SVM called the kernel [17]. Four kernels can be used for SVM such as linear kernel, polynomial kernel, radial bias function kernel, and sigmoid kernel [18].

III. METHODOLOGY

In this study, we classify toxic comments with four different categories such as pornography, defamation, hate speech, and radicalism by using the SVM method. We also run several scenarios with the goal is to obtain the best performance from overall scenarios. The scenario focused on comparing the preprocessing stages, which are stemming and stopwords removal, the implementation of Chi Square as feature selection, and the use of SVM kernels. This research includes five main steps: preprocessing, feature extraction using TF-IDF, feature selection using Chi Square, classification using SVM, and evaluation using $F1-Score$ as the metric. The flow of system shown in Figure 1 and all the detail of each step is described as follows.

A. Dataset

The dataset that we used for training and testing data is publicly available on Github conducted by [19], which collected through Instagram, Twitter, and Kaskus. The dataset was scraped and annotated manually by Ahmad Izzan, Christian Wibisono, and Ilham Firdausi Putra, and it contains of 7,773 data. The amount of data in each label and the number of multiple labels shown in Figure 2 and Figure 3, respectively.

The data is divided into four toxicity categories, including pornography, defamation, hate speech, and radicalism as shown in Table I. Each data may contain more than one toxicity categories or not possess any toxicity at all (non-toxic comment). All the labels we used in this study are explainedas follows. All the examples given below are translated from Bahasa Indonesia to English.

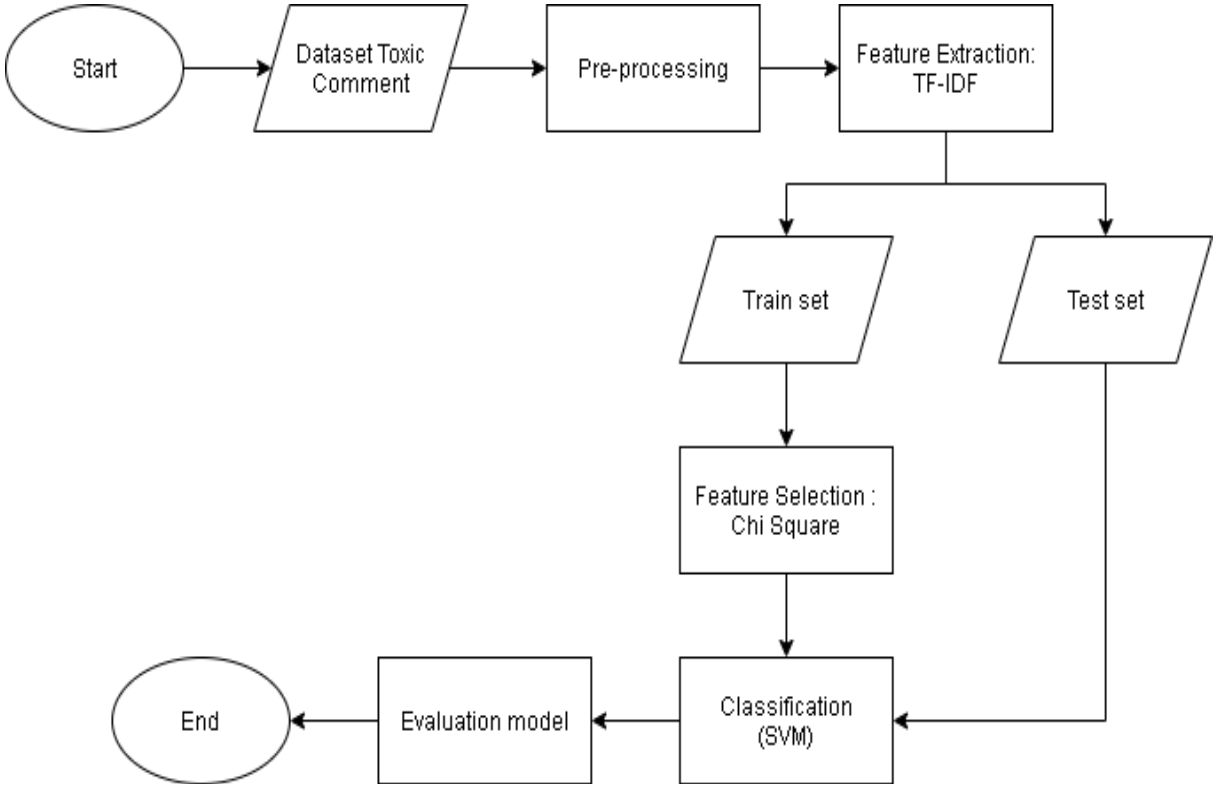

Fig. 1. The process of toxic comment classification

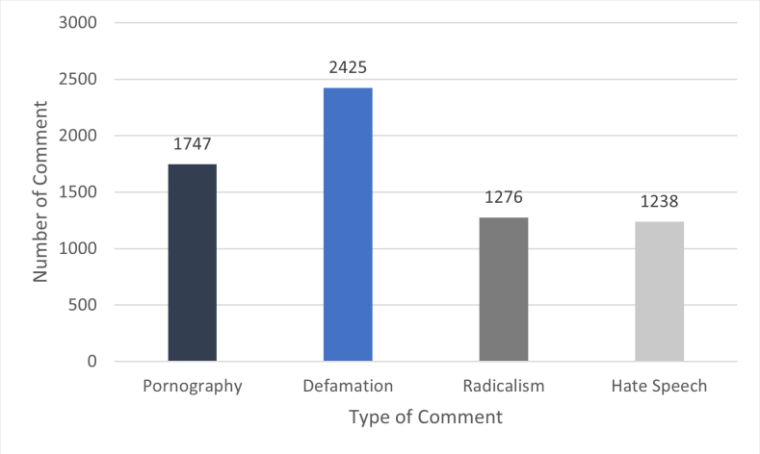

Fig. 2. Number of comment for each label in dataset

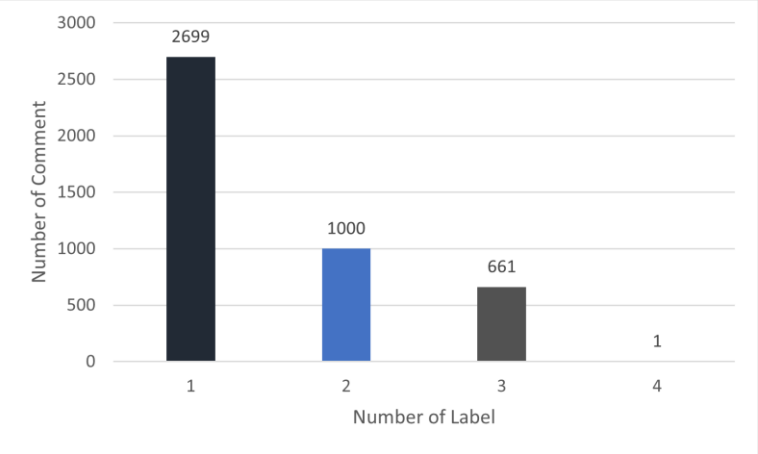

Fig. 3. Number of comment with multiple labels in dataset

TABLE I
THE EXAMPLE OF DATA IN DATASET

Comment	Pornography	Hate Speech	Defamation	Radicalism
you are the one who did sodomy but the punishment that hits not only you, shame on you	Yes	No	Yes	No
yes, my brother tried to comment on one of his posts but it was block. Please check if the account has spread hatred	No	No	No	No
can the government not be massively stupid?	No	No	Yes	No

1) **Pornography.** This label indicates a person or group of people who harass others in text form. An example of this label is “you are the one who did sodomy but the punishment that hits not only you, shame on you”.
2) **Defamation.** This label indicates someone who attacks the honor or reputation of someone by

accusing something to make it publish to the public. An example of this label is “Hey, you look like a crackhead to me”.

3) **Hate speech.** This label indicates a person or group of people who spread hate speech towards a particular ethnicity, race, and religion. An example of this label is “I don’t like you because you are Chinese”.
4) **Radicalism.** This label indicates a person or group of people who have the goal of electoral reform, such as redistribution of property rights, abolition of titles, and usually closely related to liberalism. An example of this label is “I support radicalism. Even religious radicalism. Because radical means thinking down to the essence of religion. What is the essence of religion? Benefit, kindness, humanity.”.

We choose this dataset due to the novelty of categories. Most existing datasets were established with common toxicity categories only, such as distinguishing between toxic and non-toxic comments, and indicate the category and its level of toxicity [8], [20], [9], [21], especially in the Bahasa Indonesia dataset. According to [22], toxicity has many different sub-categories such as sexism, racism, pornography, etc. Therefore, we conducted the study based on this dataset.

B. Preprocessing

In this study, there are five preprocessing stages which are data cleaning, translate text-based emojis, word normalization, stemming, and stopwords removal. The following is a brief explanation of each stage.

1) ***Data Cleaning***
In this stage we eliminate punctuation, numbers, excess space, remove URLs, formatting on Twitter, Instagram, and Kaskus. With the example “bisa coba di buka tautannya http://bit.ly/2OjnVrJÂ” (you can try to open the link http://bit.ly/2OjnVrJÂ) transformed into "bisa coba di buka tautannya" (you can try to open the link).

2) ***Translate text-based emojis***
In this stage we convert all emojis in the dataset to text by using the emoji dataset from [19] witha total of 116 emojis. With the example “*Ada apa dengan negeri ini?* :(” (what’s wrong with this country :() transformed into "*Ada apa dengan negeri ini Sedih*?" (what’s wrong with this country? Sad).

3) ***Word normalization***
In this stage of pre-processing, we eliminate repetitive characters such as “yeeees” to “yes”. Subsequently, we also change non-standard words to standard by using a dataset that comes from [19] and consists of 2,878 words. With the example “*Bodor banget lu jadi orang*” (How stupid you are) transformed into "*Bodoh sekali kamu jadi orang*" (How stupid you are).

4) ***Stemming***
In this stage, we change the Indonesian words in the dataset into basic words by removing prefixes, infixes, and suffixes. Stemming is done using PySastrawi. With the example “*dia memikul beban yang teralu berat*”(He carried too much problem) transformed into "*dia pikul beban yang lalu berat*"(He carried too much problem).

5) ***Stopwords Removal***
In this stage we removed the stopwords in Indonesian and English in the dataset using PySastrawi and Gensim, respectively. With the example “*dia memikul beban yang terlalu berat*” (He carriedtoo much problem) transformed into "*memikul beban yang terlalu berat*" (carried too much problem).

Total features (unique words) before the preprocessing stage are 40,697. After applying data cleaning, translate text-based emojis, word normalization, and stemming reduced to 29,471 features. Furthermore, after the stopwords removal process was carried out, the number of features reduced to 22,906.

C. Feature Extraction

In this study, we used TF-IDF as feature selection which is commonly used in text classification tasks such as in [23], [24], and [25]. Feature extraction process by using TF-IDF can be seen in Figure 4.

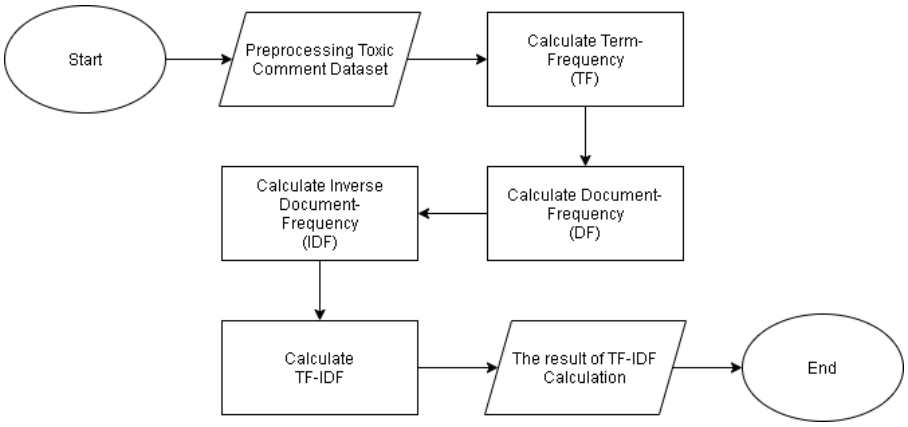

Fig. 4. Feature extraction using TF-IDF

D. Feature Selection

In this study, we used Chi Square as feature selection which is often used in multilabel classifications as in [5], [7] and [26]. Feature selection flow is showed in figure 4. Comparison is made by taking the best features of 20%, 40%, 60%, 80%, and without using Chi Square from overall dataset and implement these scenarios for each kernel. The best feature is selected based on the results of Chi Square calculation, which sorted from the highest to the lowest. Furthermore, we took the only feature that according to the proportion of scenarios. The purposes are to explore and obtain the best performance.

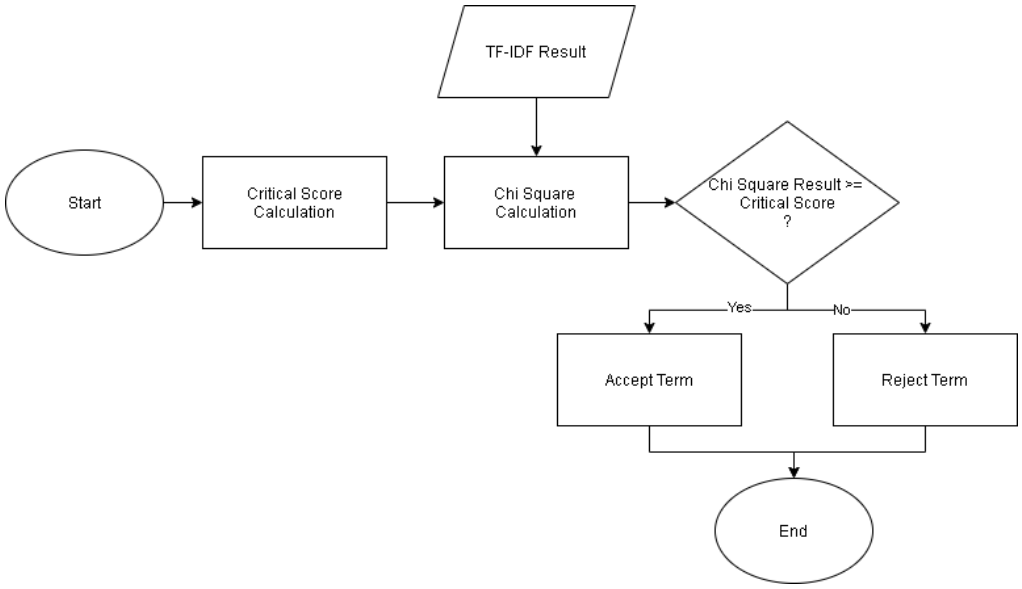

Fig. 5. Feature selection using Chi Square

E. Classification

In this study, we split the dataset randomly into two different data subsets, which are the training dataset and the testing dataset. The proportion of each dataset is 80% for train data and 20% for test data. Since the dataset was splitting randomly, we performed 10 repetitions of experiments for each scenario and took the average based on the experiment results as the final output. Since the task was multilabel classification, all of the data perhaps contain no label and one or more labels. Based on theprior research, we choose SVM due to its best performance [5]. Furthermore, we use One-vs-Rest (OvR) classifier that train one classifier per class based on a total of N different binary classifiers [27].

The kernels that we used in this study are linear, RBF, and Sigmoid. We implemented these models using a machine learning framework, namely Scikit-learn and executed the models in a Python language environment. In this experiment, we used the default parameter from Scikit-learn for all of each kernel. All the parameters are shown in Table II.

TABLE II
PARAMETER OF EACH KERNEL

Kernel	c	$gamma$	$coef0$
Linear	1.0	-	-
RBF	1.0	scale	-
Sigmoid	1.0	scale	0.0

Where c is the regularization parameter, $gamma$ is the Kernel coefficient for RBF and sigmoid only with the formula of *scale* is shown in Equation 4, and $coef0$ is an independent term in kernel function for sigmoid only.

$$\frac{1}{totalfeatures \times varianceoffeatures} \tag{4}$$

F. Evaluation

In order to be able to evaluate the performance of the model, we used $F1-Score$ as the metric. This metric computed for the minority class as the harmonic mean between precision and recall that shownin Equation 5, 6, and 7.

$$Precision = \frac{TP}{(TP + FP)} \tag{5}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{6}$$

$$F1 - Score = \frac{2 \times Precission \times Recall}{Precission + Recall} \tag{7}$$

where TP represents the number of data that are correctly classified in the minority class, FP the number of data that are incorrectly classified in the minority class, FN the number of data that belongto the minority class. However, the data is incorrectly classified in the majority class.

IV. EXPERIMENT RESULTS AND DISCUSSION

In this research, we conducted two experiment scenarios. In the first scenario, we compared several SVM kernels, including linear, sigmoid, and RBF. Subsequently, used no Chi Square and Chi Square withthe proportions 20%, 40%, 60%, and 80%. And also implemented stemming and stopwords removal. The second

scenario is the same as the first scenario. However, without implemented stemming and stopwords removal. The experiment result for each scenario can be seen in Table III and Table IV.

TABLE III
$F1-Score$ RESULTS FOR MODEL WITHOUT STEMMING AND STOPWORDS REMOVAL

Feature Selection	SVM Kernel	Training Time (s)	$F1-Score$
Chi Square (k=20%, 5,894)	Linear	16.239	75.61%
	Sigmoid	15.794	74.51%
	RBF	24.356	74.34%
Chi Square (k=40%, 11,788)	Linear	21.134	75.57%
	Sigmoid	20.245	75.34%
	RBF	30.877	72.90%
Chi Square (k=60%, 17,683)	Linear	25.104	75.76%
	Sigmoid	23.478	75,38%
	RBF	37.915	73.65%
Chi Square (k=80%, 23,577)	Linear	27.212	75.94%
	Sigmoid	25.269	74.88%
	RBF	46.196	74.22%
Non Chi Square (29,471)	Linear	28.024	76.00%
	Sigmoid	25.710	74.94%
	RBF	47.333	74.17%

TABLE IV
$F1-Score$ RESULTS FOR MODEL WITH STEMMING AND STOPWORDS REMOVAL

Feature Selection	SVM Kernel	Training Time (s)	$F1-Score$
Chi Square (k=20%, 4,581)	Linear	13.197	76.55%
	Sigmoid	13.081	74.33%
	RBF	20.951	76.47%
Chi Square (k=40%, 9,162)	Linear	17.295	75.75%
	Sigmoid	16.960	75.10%
	RBF	26.327	74.23%
Chi Square (k=60%, 13,744)	Linear	20.151	76.14%
	Sigmoid	19.225	75,56%
	RBF	31.571	74.58%
Chi Square (k=80%, 18,325)	Linear	21.920	76.35%
	Sigmoid	20.612	75.34%
	RBF	37.886	74.37%
Non-Chi Square (22,906)	Linear	22.022	76.57%
	Sigmoid	20.749	75.24%
	RBF	39.160	74.25%

Based on Table III and IV, with or without implemented stemming and stopwords obtain the fastest classification time by using the Sigmoid model. Meanwhile, the best $F1-Score$ is using the Linear model due to high dimensional input space that up to 29,000 features because it is not a mutually exclusive problem. Moreover, the Linear model has overfitting protection, so it is not dependent on a large number of features.

RBF kernel obtains the lowest $F1-Score$ and the slowest time in classification due to itacts as a prior that selects out smooth solutions that are not necessary for text classification.

The best $F1-Score$ is equal to 76.57% with 22.022 seconds of training time by executing the second scenario that implemented stemming and stopwords removal due to the non-relevant words that were removed from the dataset, including prepositions, pronouns, adverbs, etc. and transformed each word into its stem. Subsequently, the model obtained the best $F1-Score$ without implementing Chi Square, which means all the words considered as the input. The differences of $F1-Score$ between both scenarios are only 0.57%. However, the difference in execution time is 6.002 seconds.

Preprocessing stage does not significantly affect the performance of our model due to how stemming and stopwords removal worked and the amount of noise data. Stemming transforms a word into itsstem, which improves the performance due to reducing the overlap in the dataset and stopwords removal could overcome unnecessary words, which decrease the vector space and reduces the text size. Still,due to the amount of noise data such as typography, non-standard words, and an abbreviation, it doesnot significantly differ between the stages, with or without stemming and stopwords removal. Moreover, selecting features using Chi Square means reducing the redundant data so that consuming less time of training. However, implementing Chi Square does not perform the best performance due to the noise of data that leads to misunderstanding the context of the text.

For further analysis, we perform error analysis based on the result of prediction using confusion matrix to find out the remaining problem which probably causes the misclassification of data [28]. Figure 6 showsthat the majority type of error is false negative induced by 11% in defamation followed by 9.24% in hate speech, 6.42% in radicalism, and 5% in pornography. False negative become the majority type of error due to the enormous amount of unbalanced data as shown in Figure 3. Because of the massive amountof unbalanced data, the majority class will obtain a better classification performance than the minority class, which impact to give a large amount of false negative error [29].

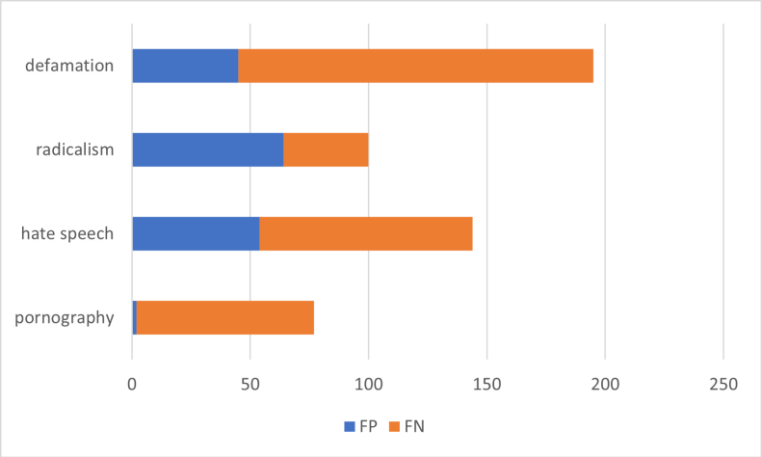

Fig. 6. Error analysis based on the number of False Positive (FP) and False Negative (FN) from each labels

We observe that plenty of misclassification occur in defamation was caused by the lack of specific keywords that represent the context of a label and perpetually conveyed implicitly. Whereas slightly number of errors in pornography was caused by the keywords that often occur in a typical comment and mostly followed by swear words, becoming easier to characterize. Besides, we do manual inspection in general on the predicted dataset by using 500 random samples from 1,555 total data. We observe three main problems in predicted data referred to previous research [30].

1) ***Out-of-vocabulary (OOV) words****:* We encountered a similar problem as the previous research where OOV is one of the main problems in this task due to the existence of words in train set. It is attributedto human writing styles such as abbreviation, typography, slang words, etc. This problem may decrease the performance due to misinterpretation of the context of a sentence.

2) ***Multi-word expressions****:* There are several multi-word expressions that occur in the dataset, such as metaphorical expression (*Do not trust our government, they words are venomous snakes*), verbal idioms(*you are only a girl scout and silence is gold darling*), and quotations (*"you may learn the meaning of infidel", "Try talking about the bomb on the plane". None of these words are making sense, think before you speak).* It was becoming a problem since our models could not recognize these types of expressions.

3) ***Usage of Different Languages****:* As we know, Indonesia has a variety of regional languages. The vocabulary among region was completely different, especially on the different island. Therefore, a lot of comments is written in several languages. The problem may lead to misclassification of data due to the inability of models in recognizing and translating the languages. For that, we add this new type of error.

V. CONCLUSION

The ease of using social media leads the irresponsible user to do unethical things, including spread hatred, defamation, radicalism, pornography which are categorized as toxic comments. We classified the toxic comment into its toxicity category using data from Twitter, Kaskus, and Instagram. We did two experimental scenarios, including comparing several SVM kernels, including linear, sigmoid, and RBF. Subsequently, implemented Chi Square as the feature selection with the proportions 20%, 40%, 60%, and 80%. And also implemented stemming and stopwords removal. The following scenario is the same as the first scenario but without implemented stemming and stopwords removal. By running these scenarios, the best model was obtained using SVM as the model with a linear kernel, without using Chi Square, and using stemming and stopwords removal with the $F1-Score$ equal to 76.57%. The differences of $F1-Score$ between both scenarios are only 0.57%. However, the difference in execution time is 6.002 seconds.

Comparing to the previous work [3] as mentioned in section II, by applying the method and scenario to our dataset, we obtained 76,57% of $F1-Score$. In this dataset, applying stemming and stopwords removal obtained the highest $F1-Score$ while in the previous research, the highest $F1-Score$ was obtained by not applying those scenarios. The performance in our dataset is 1,69% better than the previous work. Besides, we conducted more exploration by applying Chi Square as the feature selection based on the previous work that has been proposed by [5]. Unfortunately, applying Chi Square on our dataset does not give the $F1-Score$ result as we expected. However, the execution is less time-consuming.

To enhance the performance of the model, several suggestions can be done. First, since an imbalance of the dataset, we may use several methods and algorithms to balance the data, such as manual annotation, text augmentation, or using several balancing algorithms such as Synthetic Minority Oversampling Technique (SMOTE), WEMOTE/CWEMOTE, Penalize, etc. Subsequently, to enrich the knowledge of the model in recognizing the words input, we can add word embedding such as word2vec, FastText, etc.

RFERENCES

[1] D. N. Lapedes, McGraw-Hill dictionary of scientific and technical terms. 1974.

[2] S. Kemp, “Digital 2020 indonesia,” 2020.

[3] K. M. Hana, S. Al Faraby, A. Bramantoro, et al., “Multi-label classification of indonesian hate speech on twitter using support vector machines,” in 2020 International Conference on Data Science and Its Applications (ICoDSA), pp. 1–7, IEEE, 2020.

[4] S. Malmasi and M. Zampieri, “Detecting hate speech in social media,” arXiv preprint arXiv:1712.06427, 2017.

[5] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, “Feature selection using an improved chi-square for arabic text classification,” Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 2, pp. 225–231, 2020.

[6] A. A. Sagar and J. S. Kiran, “Toxic comment classification using natural language processing,” 2008.

[7] A. Y. Taha and S. Tiun, "Binary relevance (br) method classifier of multi-label classification for arabic text.," Journal of Theoretical & Applied Information Technology, vol. 84, no. 3, 2016.

[8] W. Abbas, Toxic Comment Classification of Roman Urdu Text. PhD thesis, Department of Computer Science, COMSATS University Islamabad, Lahore Campus, 2019.

[9] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in indonesian twitter," in Proceedings of the Third Workshop on Abusive Language Online, pp. 46–57, 2019.

[10] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," International Journal of Data Warehousing and Mining (IJDWM), vol. 3, no. 3, pp. 1–13, 2007.

[11] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou, "Multi-label learning with weak label," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 24, 2010.

[12] A. Wieczorkowska, P. Synak, and Z. W. Ras´, "Multi-label classification of emotions in music," in Intelligent Information Processing and Web Mining, pp. 307–315, Springer, 2006.

[13] J. Ye, X. Jing, and J. Li, "Sentiment analysis using modified lda," in International conference on signal and information processing, networking and computers, pp. 205–212, Springer, 2017.

[14] I. Yulietha, S. Faraby, and A. Adiwijaya, "Klasifikasi sentimen review film menggunakan algoritma support vector machine,"eProceedings of Engineering, vol. 4, no. 3, 2017.

[15] H. Syahputra, L. Basyar, and A. Tamba, "Setiment analysis of public opinion on the go-jek indonesia through twitter using algorithm support vector machine," in Journal of Physics: Conference Series, vol. 1462, p. 012063, IOP Publishing, 2020.

[16] N. Cristianini, J. Shawe-Taylor, et al., An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[17] G. M. Fung, O. L. Mangasarian, and J. W. Shavlik, "Knowledge-based nonlinear kernel classifiers," in Learning Theory and Kernel Machines, pp. 102–113, Springer, 2003.

[18] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al., "A practical guide to support vector classification," 2003.

[19] A. Izzan, C. Wibisono, and I. F. Putra, "Indonesian social media post toxicity classification." https://github.com/ahmadizzan/ netifier, 2018.

[20] S. Zaheri, J. Leath, and D. Stroud, "Toxic comment classification," SMU Data Science Review, vol. 3, no. 1, p. 13, 2020.

[21] 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, Springer, 2021.

[22] A. L. Association, "Hate speech and hate crime," 2017.

[23] L.-P. Jing, H.-K. Huang, and H.-B. Shi, "Improved feature selection approach tfidf in text mining," in Proceedings. International Conference on Machine Learning and Cybernetics, vol. 2, pp. 944–946, IEEE, 2002.

[24] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, et al., "News article text classification in indonesian language," Procedia Computer Science, vol. 116, pp. 137–143, 2017.

[25] B. Trstenjak, S. Mikac, and D. Donko, "Knn with tf-idf based framework for text categorization," Procedia Engineering, vol. 69, pp. 1356–1364, 2014.

[26] L. Chekina, L. Rokach, and B. Shapira, "Meta-learning for selecting a multi-label classification algorithm," in 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 220–227, IEEE, 2011.

[27] S. Chatterjee, P. G. Jose, and D. Datta, "Text classification using svm enhanced by multithreading and cuda.," International Journal of Modern Education & Computer Science, vol. 11, no. 1, 2019.

[28] T. Fawcett, "An introduction to roc analysis: Pattern recognition letter, v. 27," 2006.

[29] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 4, pp. 42–47, 2012.

[30] B. van Aken, J. Risch, R. Krestel, and A. Löser, "Challenges for toxic comment classification: An in-depth error analysis,"arXiv preprint arXiv:1809.07572, 2018.