An Explainable AI Model for Hate Speech Detection on Indonesian Twitter

Muhammad Amien Ibrahim[1], Samsul Arifin[2*], I Gusti Agung Anom Yudistira[3], Rinda Nariswari[4],
Abdul Azis Abdillah[5], Nerru Pranuta Murnaka[6], and Puguh Wahyu Prasetyo[7]
[1]Computer Science Department, School of Computer Science, Bina Nusantara University
Jakarta 11480, Indonesia
[2−4]Statistics Department, School of Computer Science, Bina Nusantara University
Jakarta 11480, Indonesia
[5]Department of Mechanical Engineering, Politeknik Negeri Jakarta
Depok 16425, Indonesia
[6]Department of Mathematics Education, STKIP Surya
Tangerang 15334, Indonesia
[7]Mathematics Education Department, Universitas Ahmad Dahlan
Daerah Istimewa Yogyakarta 55166, Indonesia
Email: [1]muhammad.ibrahim1@binus.edu, [2]samsul.arifin@binus.edu, [3]i.yudistira@binus.ac.id,
[4]rinda.nariswari@binus.ac.id, [5]abdul.azis.a@mesin.pnj.ac.id, [6]nerru.pranuta@stkipsurya.ac.id,
[7]puguh.prasetyo@pmat.uad.ac.id

***Abstract*—To avoid citizen disputes, hate speech on social media, such as Twitter, must be automatically detected. The current research in Indonesian Twitter focuses on developing better hate speech detection models. However, there is limited study on the explainability aspects of hate speech detection. The research aims to explain issues that previous researchers have not detailed and attempt to answer the shortcomings of previous researchers. There are 13,169 tweets in the dataset with labels like "hate speech" and "abusive language". The dataset also provides binary labels on whether hate speech is directed to individual, group, religion, race, physical disability, and gender. In the research, classification is performed by using traditional machine learning models, and the predictions are evaluated using an Explainable AI model, such as Local Interpretable Model-Agnostic Explanations (LIME), to allow users to comprehend why a tweet is regarded as a hateful message. Moreover, models that perform well in classification perceive incorrect words as contributing to hate speech. As a result, such models are unsuitable for deployment in the real world. In the investigation, the combination of XGBoost and logical LIME explanations produces the most logical results. The use of the Explainable AI model highlights the importance of choosing the ideal model while maintaining users' trust in the deployed model.**

***Index Terms*—Artificial Intelligence Model, Hate Speech, Indonesian Twitter**

Received: March 16, 2022; received in revised form: April 19, 2022; accepted: April 19, 2022; available online: Sep. 14, 2022.
*Corresponding Author

I. Introduction

HATE speech is directed at a person or group that contains hatred based on something about that person or organization. Ethnicity, religion, disability, gender, and sexual orientation are all frequently exploited as justifications for hatred. Hate speech propagation is an extremely harmful behavior that can result in discrimination, social strife, and human genocide as governments, businesses, and researchers have all made considerable investments in countermeasures [1]. For example, the Tutsi ethnic genocide in Rwanda in 1994 was one of the most brutal genocides triggered by the act of propagating hate speech [2]. The tragedy was caused by hate speech spread by some parties, which claimed that the Tutsi ethnic group was the source of increased political, economic, and social pressure [3].

Hate speech is frequently accompanied by abusive language in everyday life, particularly on social media [4]. Then, abusive language is an utterance that incorporates abusive words/phrases and is delivered vocally or in writing to the interlocutor (individuals or groups) [5, 6]. In Indonesia, abusive words are usually derived from an unpleasant condition, such as mental illness, sexual deviation, physical disability, a lack of modernization, a condition where someone lacks etiquette, conditions that are not allowed by religion, and other conditions related to unfortunate circumstances, animals with a bad characteristic, disgusting and forbidden things in a certain religion, astral beings that frequently interfere with human life, and a dirty and filthy environment [7]. Because of the use of abusive words/phrases that elicit emotions, hate speech that involves abusive words/phrases frequently accelerates the onset of social conflict [4, 6].

In Indonesia, offensive statements intended to curse someone (spreading hate speech) are separated into three categories: words, phrases, and clauses. Although harsh language is sometimes used as a joke (not to insult someone), its usage on social media can generate conflict due to misunderstandings among netizens. Furthermore, children may be exposed to the inappropriate language for their age as a result of harsh language seen on social media [8]. To avoid disputes between people and children who learn hate speech and improper language from the social media they use, hate speech and abusive language on social media must be caught. Some researchers have recently investigated hate speech detection [9] and abusive language detection [10].

Hate speech has a specific goal, category, and degree [11]. Hate speeches fall under various categories, including ethnicity, religion, race, sexual orientation, and others, and are directed against a specific individual or group with a high level of hatred [12]. However, no research on hate speech identification and explanation has been undertaken simultaneously, according to the literature review. Many hate speech identification studies focus solely on determining whether a text is hate speech or not [13]. For example, previous research detects the amount of hate speech. It categorized Italian Facebook posts and comments into three categories: no hate speech, mild hate speech, and strong hate speech. However, it does not explain why a comment is regarded as hate speech [10].

Many studies in abusive language identification only determine whether a text is an abusive language or not. It is similar to hate speech detection research. An example is a study identifying hate speech and abusive language on Indonesian Twitter. It categorizes Indonesian tweets into three categories: no hate speech, abusive but not hate speech, and abusive and hate speech. However, like previous studies on hate speech and abusive language identification, it does not explain why a tweet is deemed a hateful message [14].

Twitter is one of the social media platforms in Indonesia that is frequently used to promote hate speech, so the researchers choose it as the dataset. The research is a text classification issue, in which a tweet can be classified as no hate speech or hate speech. The researchers utilize a machine learning approach with various classifiers to detect multi-label hate speech and abusive language. Logistic Regression, Multinomial Naive Bayes, Random Forest Decision Tree (RFDT), and XGBoost are the classifiers used [15]. Based on past research, these classifiers are algorithms that can detect hate speech in Indonesian with reasonable accuracy. Term Frequency-Inverse Document Frequency (TF-IDF) is the text classification characteristic used. To evaluate the suggested technique, the researchers utilize accuracy. Furthermore, a model-interpretability approach, such as Local Interpretable Model-Agnostic Explanations (LIME), is used to provide an explanation [16]. Generally, accuracy metrics are used to evaluate classification models. On the other hand, real-world data is frequently different since collected and annotated data may contain bias. In some cases, the accuracy metric may not represent the main objective of building the text classification model. Therefore, in addition to such metrics, evaluating individual predictions like the LIME technique can provide an alternative solution [3].

LIME, as an explainable model, ensures that the model is implementable by revealing its inner functions. This technique can explain the complex classification model by providing explanations through the less complex model. For instance, a sample tweet is predicted using a complex classification model, and LIME will create a less complex model, such as linear regression, to explain how this sample tweet is classified to a certain class. The coefficients in the linear regression will provide information on how much they will affect the classification output. LIME is a model agnostic technique that may be used with any machine learning model [16]. It provides some flexibility in interpreting and explaining model prediction without limiting the option of what classification models to develop. Furthermore, the ability to provide explanations through the LIME technique will assist decision-makers in gaining trust in the chosen model, bringing it one step closer to deployment in the real world [17].

The Explainable AI model has been utilized in several previous studies. First, an Explainable AI model is developed to explain xenophobic tweets to aid decision-makers in preventing acts of violence [18]. It is also attempted to counter toxic comments on social media, such as YouTube comments, by developing classification models and choosing the final model using LIME explanations to eliminate biased models [17]. Another hate speech detection is developed using a new benchmark English tweets dataset, and LIME explanations evaluate the bias and interpretability [3]. The new benchmark English tweets dataset compares and evaluates hate speech words highlighted by LIME and hate speech words highlighted by human annotators. Another classification task is carried out in different domains [19]. The LIME technique is employed in legal document classification to provide explanations for how documents are classified into different categories, allowing lawyers to inspect documents more efficiently.

Previous research has some limitations, such as competing to make models with good performance but does not explain why a prediction is classified into a particular class [5]. In some countries, expressing hateful messages towards an individual or group can be prosecuted. Thus, the prediction results from a hate speech detection model should not be simply trusted, considering there are legal consequences. Moreover, identifying hate speech and providing a logical reason is critical in assisting authorities in prioritizing hate speech instances that need to be addressed immediately.

The value of the research is that it explains everything that prior researchers have not explained in depth. The research attempt to address past researchers' limitation. The research investigates the identification and explanation of hate speech on Indonesian Twitter. Explaining a prediction implies identifying words that give a qualitative understanding of their relevance to the model's prediction. In this example, an explanation is a small group of words that can contribute to or contradict the prediction.

II. Research Method

The researchers use a prior study's hate speech and abusive language datasets on Twitter. There are 13,169 tweets in the dataset with labels such as “hate speech” and “abusive language”. The dataset also provides binary labels on whether hate speech is directed to individual, group, religion, race, physical disability, and gender. In addition, the dataset is complemented with binary labels that specify the degree of hate speech (weak, moderate, and strong). Hate speech is when someone expresses animosity toward another individual or group based on ethnicity, religion, disability, gender, or sexual orientation. On the other hand, abusive words are typically derived from a situation, such as mental illness, sexual deviation, physical impairment, and other unfavorable conditions that frequently accelerate the formation of social conflict due to the usage of abusive derogatory remarks that provoke emotions [5]. There are 5,860 tweets labeled as hate speech and 2,266 tweets as abusive language among the total number of tweets. Furthermore, 4,143 tweets have been flagged as both hate speech and abusive language. Because hate speech is frequently aimed against a specific category, such as an individual or a group, two more labels are added to this category [7].

TABLE I

An Example of a Tweet Flagged as a Weak Hate Speech and Abusive Language Targeting Individuals Based on Gender.

Column	Value
Tweet	USER USER dasar duo homo sarap hadehh pusing pala bebi (you are a crazy gay duo giving me headache)
HS	Yes
Abusive	Yes
HS_Individual	Yes
HS_Group	No
HS_Religion	No
HS_Race	No
HS_Physical	No
HS_Gender	Yes
HS_Other	No
HS_Weak	Yes
HS_Moderate	No

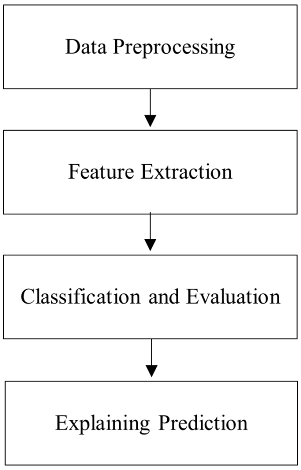
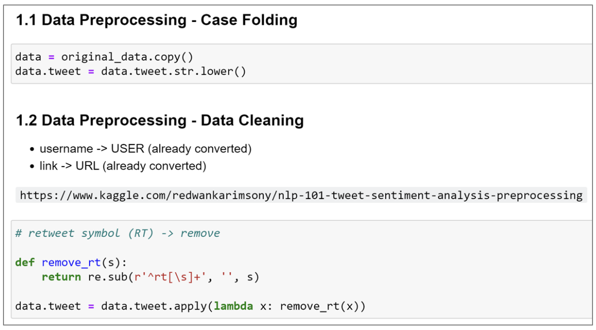


Fig. 1. Flowchart of steps performed in the experiment.

Then, a set of more granular categorizations of the group is also included as labels like religion, race, physical disability, and gender, as mentioned before. Table I is an example of a tweet with its labels. The tweet row specifies the message posted by the user. The other rows underneath specify whether the post is included in such category. For example, Table I shows that a tweet is considered abusive and has hate speech directed at an individual and gender and a weak level of hatred. HS stands for hate speech.

The experiment makes use of tweets as well as the hate speech labels that go along with them. Because tweets are textual data, multiple text processing processes are used. The features from the tweets are retrieved and utilized as input for classification algorithms. A model-agnostic Explainable AI technique is employed in the final phase to explain why a tweet is labeled as hate speech or not hate speech. Figure 1



1.1 Data Preprocessing - Case Folding

```
data = original_data.copy()
data.tweet = data.tweet.str.lower()
```

1.2 Data Preprocessing - Data Cleaning

- username -> USER (already converted)
- link -> URL (already converted)

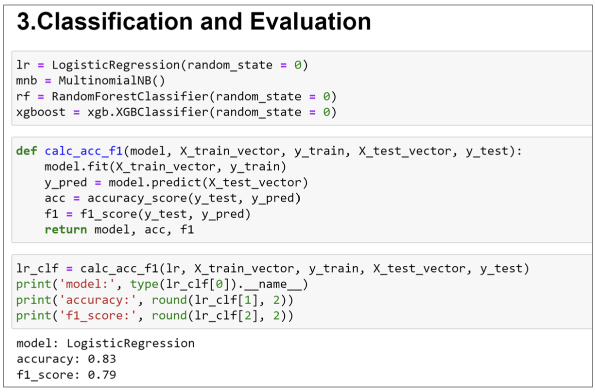
```
https://www.kaggle.com/redwankarimsony/nlp-101-tweet-sentiment-analysis-preprocessing
```

```
# retweet symbol (RT) -> remove

def remove_rt(s):
    return re.sub(r'^rt[\s]+', '', s)

data.tweet = data.tweet.apply(lambda x: remove_rt(x))
```

Fig. 2. Python code for data preprocessing.



3.Classification and Evaluation

```
lr = LogisticRegression(random_state = 0)
mnb = MultinomialNB()
rf = RandomForestClassifier(random_state = 0)
xgboost = xgb.XGBClassifier(random_state = 0)
```

```
def calc_acc_f1(model, X_train_vector, y_train, X_test_vector, y_test):
    model.fit(X_train_vector, y_train)
    y_pred = model.predict(X_test_vector)
    acc = accuracy_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    return model, acc, f1
```

```
lr_clf = calc_acc_f1(lr, X_train_vector, y_train, X_test_vector, y_test)
print('model:', type(lr_clf[0]).__name__)
print('accuracy:', round(lr_clf[1], 2))
print('f1_score:', round(lr_clf[2], 2))
```

```
model: LogisticRegression
accuracy: 0.83
f1_score: 0.79
```

Fig. 3. Python code for classification and evaluation.

shows the flowchart describing the steps taken in the experiment. Then, in Figs. 2–5, the researchers create the full Python code, which becomes the source of the main results of this article. Those codes can be accessed at the following link: https://github.com/amnibrahim/xai_hatespeech_indonesian_tweet.

Figure 2 highlights the Python code for the data preprocessing. The process starts with data preprocessing. It is followed by extracting features from cleaned data. Then, models are trained and evaluated. Finally, an explainable AI model is developed to provide explanation to many predictions. Moreover, the Python code for the classification and evaluation is shown in Fig. 3 [20].

As in a general data science pipeline, once the data has been preprocessed, several classification models are chosen. In the research, the classification models are Logistic Regression, Multinomial Naive Bayes, Random Forest, and XGBoost. These classification models are considered the black box models. A black box model is a model that receives input and generates output without revealing inner workings and limiting users from understanding its inner functions [16]. In the context of Explainable AI, these black box models are the models that the researchers want to learn more about through explanations. Then, the preprocessed data are predicted using the black box models.

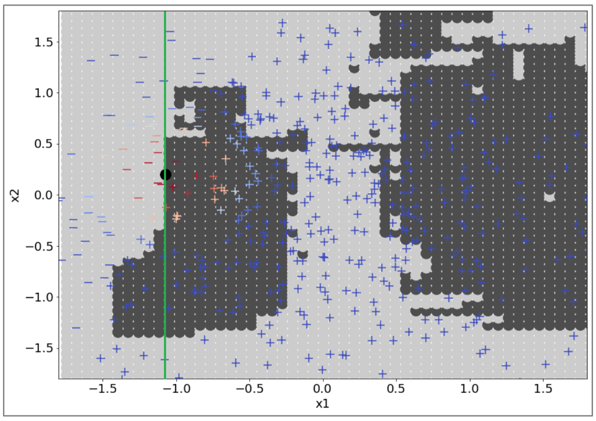


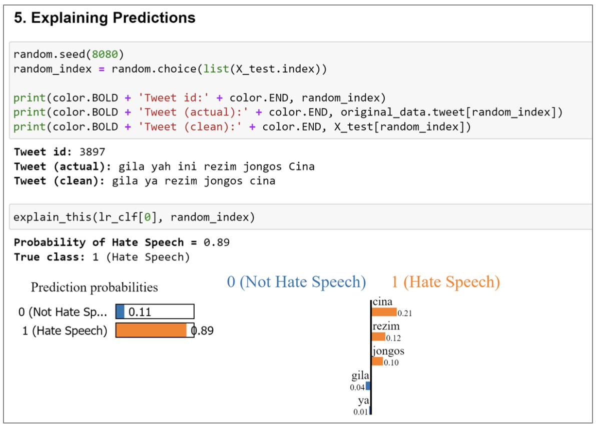
Fig. 4. Decision boundary of a complex and less complex model.

The following step is to pick one sample from the dataset to be predicted. The black dot in Fig. 4 depicts the selected single sample that the output prediction will be explained using the LIME technique. Following that, random data are generated in the area around the selected single sample since the researchers only concern with the immediate vicinity of the selected single sample. These new data are generated by perturbations.

As an example, the decision boundary from a black box model can be illustrated by the dark gray and light gray in Fig. 4. The simpler model, such as the linear regression model, may find it hard to simulate. The decision boundary of the linear regression model is shown as a green line in Fig. 4. Eventually, the fitted linear regression model's coefficients will allow the selected single sample to be explained.

By utilizing black box models, these newly generated random data are classified, and the output prediction is obtained. It produces a new dataset of the newly generated random data with their predicted label. In this scenario, there are two labels: hate speech and non-hate speech. Then, a weight value is assigned to each newly generated random data. The weights are determined by how close the newly generated random data is to the selected single sample. Fig. 5 indicates that the newly generated random data are colored red since they are closer to the selected single sample. Meanwhile, the others are colored in blue since they are far from the selected single sample.

A simpler and more interpretable Explainable AI model is then developed. The Explainable AI model in the research is linear regression. The newly generated random data and its label are used to fit the linear regression. The weights are utilized in the loss function calculation to optimize a linear regression model. As a reminder, the weights are calculated by computing the



5. Explaining Predictions

```
random.seed(8080)
random_index = random.choice(list(X_test.index))

print(color.BOLD + 'Tweet id:' + color.END, random_index)
print(color.BOLD + 'Tweet (actual):' + color.END, original_data.tweet[random_index])
print(color.BOLD + 'Tweet (clean):' + color.END, X_test[random_index])
```

```
Tweet id: 3897
Tweet (actual): gila yah ini rezim jongos Cina
Tweet (clean): gila ya rezim jongos cina
```

```
explain_this(lr_clf[0], random_index)
```

```
Probability of Hate Speech = 0.89
True class: 1 (Hate Speech)
```

Fig. 5. Python code for explaining prediction.

distance between each newly generated random data and the single selected sample that has been chosen. As a result, the further the data is, the fewer weights it will receive. On the other hand, the closer the data are, the more weight they will receive. It emphasizes that nearby data are more important and ensures that the model is locally faithful [21]. The following Fig. 5 shows the Python code for explaining prediction.

III. Results and Discussion

Several steps of data preprocessing are carried out to achieve the best possible result. The first is case folding, which involves transforming all characters to lowercase to make them uniform. Following that, data cleaning removes excessive characters like retweet (“RT”) punctuations and emojis. The username, hashtag, numbers, and hyperlink are then converted to “user”, “hashtag”, “number”, and “URL”, respectively. Despite emojis, hashtags, numbers, and hyperlinks that may vary, those do not provide a significant contribution to classification. Therefore, these types of strings are removed or transformed into a general label. Then, the next step is text normalization, which converts non-formal terms into formal ones. In the research, text normalization is performed using a dictionary derived from the previous study [5]. Following that, stemming is applied to each tweet in the dataset to transform inflectional words into their base form. The stemming is done with the Sastrawi Library. Furthermore, eliminating stop words is performed by using the stop word in the provided list [22]. In the final step of preprocessing, tweets with a length of three words or less are excluded from the dataset to be considered representative and to prevent misclassification [17].

TABLE II
The Performance of Each Model.

Model No.	Model Name	Accuracy Score	F1-Score
1	Logistic Regression	0.83	0.79
2	Multinomial Naïve Bayes	0.83	0.79
3	Random Forest	0.82	0.79
4	XGBoost	0.83	0.79

The research uses word n-grams for feature extraction, with each word weighted by TF-IDF. In previous studies [3, 4], the TF-IDF is utilized as a baseline for feature extraction. In addition, it is found that TF-IDF provides the highest accuracy parameter across all machine learning models used in the experiments [5]. Several traditional machine learning models are utilized for classification, including Logistic Regression, Multinomial Naive Bayes, Random Forest, and XGBoost. During the training phase, the models are trained using the skit-learn library using default parameters. The models are then compared using two different methods. The first method involves using common evaluation metrics, such as accuracy and F1-score. The second method uses LIME, as investigated in previous research [17]. LIME, a model-agnostic Explainable AI technique, is one of the most well-known contributions to the simplification approach to explanations. The simplified model that is easier to implement due to its reduced complexity compared to the model it represents is explained using the simplification approach technique. LIME creates locally linear models around its predictions to explain an opaque model [16].

The experiment is carried out by training and testing several traditional machine learning models. Then, some evaluation metrics such as accuracy and F1-score are obtained. Table II shows both evaluation metrics for its corresponding model with accuracy and F1-score results that are very close.

It is challenging to choose and reject models based on standard evaluation metrics like accuracy and F1-score. As a result, several tweets are predicted. Then, their LIME explanations are compared manually, like in previous research [17]. Several hateful tweets from the testing set are randomly selected. Three randomly selected hate tweets are predicted as hate speech and are examined through LIME explanations. Table III lists the selected hateful tweets predicted as hate speech.

Next, their LIME explanations are manually compared in Table IV. It shows the LIME explanations output for each machine learning model. The green highlighted words have been identified as contributing to the non-hate speech class. The yellow highlighted words, on the other hand, are words that shift the prediction to the hate speech class. The Logistic Regression and Multinomial Naive Bayes model in Table IV

TABLE III
THREE RANDOMLY SELECTED HATE TWEETS.

No.	Tweet (raw)
1	*Gila yah ini rezim jongos Cina* (this is crazy Chinese servant regime)
2	*Ganyang PKI..!! Perang terbuka generasi PKI dg ummat islam..!!* (destroy Indonesian Communist Party..!! Open warfare between the Indonesian Communist Party generation and Muslims)
3	USER USER USER USER *Dangkal otak kamu, tidak semua yg benci suharto itu PKI.* (your brain is shallow, not everyone who hates Suharto is Indonesian Communist Party)
No.	Tweet (clean)
1	*Gila ya rezim jongos Cina* (crazy Chinese servant regime)
2	*Ganyang Partai Komunis Indonesia perang buka generasi Partai Komunis Indonesia umat Islam* (destroy Indonesian Communist Party open warfare Indonesian Communist Party Muslims)
3	*Dangkal otak benci Soeharto Partai Komunis Indonesia* (shallow brain hate Soeharto Indonesian Communist Party)

TABLE IV
THE LIME EXPLANATIONS FOR THE FIRST TWEET.

Model	Tweet
Logistic Regression	*gila ya rezim jongos cina*
Multinomial Naïve Bayes	*gila ya rezim jongos cina*
Random Forest	*gila ya rezim jongos cina*
XGBoost	*gila ya rezim jongos cina*

TABLE V
THE LIME EXPLANATIONS FOR THE SECOND TWEET.

Model	Tweet
Logistic Regression	*ganyang partai komunis indonesia perang buka generasi partai komunis indonesia umat islam*
Multinomial Naïve Bayes	*ganyang partai komunis indonesia perang buka generasi partai komunis indonesia umat islam*
Random Forest	*ganyang partai komunis indonesia perang buka generasi partai komunis indonesia umat islam*
XGBoost	*ganyang partai komunis indonesia perang buka generasi partai komunis indonesia umat islam*

TABLE VI
THE LIME EXPLANATIONS FOR THE THIRD TWEET.

Model	Tweet
Logistic Regression	*dangkal otak benci soeharto partai komunis indonesia*
Multinomial Naïve Bayes	*dangkal otak benci soeharto partai komunis indonesia*
Random Forest	*dangkal otak benci soeharto partai komunis indonesia*
XGBoost	*dangkal otak benci soeharto partai komunis indonesia*

highlights the first-word “*gila*” (crazy) as the word that the LIME model finds important in predicting non-hate speech. It is incorrect because “*gila*” is a derogatory mark often used in hate speech and abusive language. Therefore, the first two models are rejected. The second word “*ya*” is the Indonesian translation of “yes”. Thus, it is correct to flag it as either green or no flag at all. The rest of the words are correctly highlighted as yellow for all models, meaning these words are the words that shift the prediction to hate speech class.

Table V highlights the explanation for some words for the second tweet for each model. It shows that the word “*ganyang*” (destroy), “*partai*” (political party), “*komunis*” (communist), and “*Indonesia*” are highlighted as yellow in most of the models. A notable difference can be seen in the first two models of Logistic Regression and Multinomial Naive Bayes, where the word “*perang*”, meaning “war” in English, is highlighted as green. It is incorrect as the word “war” is often used in the hate speech conversation. As a result, both models are rejected.

Table VI shows the explanation of the third tweet. Three out of four models highlight the word “*Soeharto*” as green. It is the name of an Indonesian army general who was involved in a coup and became president in 1968, eventually stepping down after 32 years in power. In the Random Forest model, the word “*benci*” (hate) and “*komunis*” (communist) are also highlighted as green. In general, these three terms are used in the context of political debate on social media by various netizens with opposing political views, potentially leading to hate speech. As a result, the first three models are rejected, and the XGBoost model is chosen as the final selected model.

Utilizing such Explainable AI technique can provide an alternative evaluation method other than common metrics, such as accuracy and F1-score. Both accuracy and F1-score achieve highly similar significant results, but the highlighted words in LIME output show different results. Among the three predicted tweets used as examples, XGBoost is considered to provide the most logical explanation since the yellow highlighted words are correctly identified as hate speech words. In contrast, green highlighted words are properly recognized non-hate speech words. At this stage, the LIME explanations can provide valuable insight for the decision makers to choose which model to be deployed. Also, it allows the decision makers to trust the prediction and the model.

IV. CONCLUSION

The research studies hate speech detection in Indonesian Twitter and attempt to explain it. A dataset from a prior study is used, which includes 13,169 tweets labeled as hate speech or not hate speech. Several traditional machine learning models are trained to determine whether a tweet is classified as hate speech or not, with the models achieving a high level of accuracy overall. However, after examining the models' predictions explanations, it is discovered that XGBoost provides the most logical LIME explanations for the predictions. As a result, model interpretability techniques, such as LIME, can assist in selecting the ideal model, among others that yield great results for deployment. The model interpretability technique can also be used to give end-users with reasoning for a model's prediction.

The research has limitations. Most words in each class are tweets related to political events. Furthermore, in terms of the dataset, there is an opportunity to collect more general data. It means that hate speech on Twitter can be collected from general scenarios and not only from specific cases, but such also as tweets that relate to political events. Another opportunity for future research is to provide tags that specify whether words in a tweet are likely to contribute to certain classes like hate speech or not. These tags can be used as a ground truth which can be compared to the LIME output. As a result, a quantitative metric can be utilized to see how different the words highlighted by the ground truth and the words chosen by LIME explanations are. Overall, future research can revolve around collecting a new general dataset of hate speech on Indonesian Twitter. The dataset can also be used as a benchmark dataset for future studies.

Acknowledgement

The researchers wish to express their gratitude to the experts for their helpful and constructive ideas for improving the research. The researchers also want to thank Bina Nusantara University. The research is possible thanks to the Binus Initiative Project grant's facility, finance, and infrastructure.

References

[1] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, 2019.

[2] G. H. Stanton, "The Rwandan genocide: Why early warning failed," *Journal of African Conflicts and Peace Studies*, vol. 1, no. 2, pp. 6–25, 2009.

[3] A. A. Abdillah, A. Azwardi, S. Permana, I. Susanto, F. Zainuri, and S. Arifin, "Performance evaluation of linear discriminant analysis and support vector machines to classify cesarean section," *Eastern-European Journal of Enterprise Technologies*, vol. 5, no. 2, pp. 37–43, 2021.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017, pp. 512–515.

[5] 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*, 2019, pp. 46–57.

[6] H. Yenala, A. Jhanwar, M. K. Chinnakotla, and J. Goyal, "Deep learning for detecting inappropriate content in text," *International Journal of Data Science and Analytics*, vol. 6, no. 4, pp. 273–286, 2018.

[7] S. Arifin, I. B. Muktyas, W. F. Al Maki, and M. K. B. Mohd Aziz, "Graph coloring program of exam scheduling modeling based on Bitwise coloring algorithm using Python," *Journal of Computer Science*, vol. 18, no. 1, pp. 26–32, 2022.

[8] S. Frenda, B. Ghanem, M. Montes-Y-Gómez, and P. Rosso, "Online hate speech against women: Automatic identification of misogyny and sexism on Twitter," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4743–4752, 2019.

[9] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. Bali, Indonesia: IEEE, Oct. 28–29, 2017, pp. 233–238.

[10] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," in *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 2017, pp. 86–95.

[11] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 145–153.

[12] M. O. Ibrohim, E. Sazany, and I. Budi, "Identify abusive and offensive language in Indonesian Twitter using deep learning approach," *Journal of Physics: Conference Series*, vol. 1196, no. 1, pp. 1–6, 2019.

[13] S. Tuarob and J. L. Mitrpanont, "Automatic discovery of abusive Thai language usages in social networks," in *International Conference on Asian Digital Libraries*. Bangkok, Thailand: Springer, Nov. 13–15, 2017, pp. 267–278.

[14] S. Arifin, I. B. Muktyas, P. W. Prasetyo, and A. A. Abdillah, "Unimodular matrix and Bernoulli map on text encryption algorithm using Python," *Al-*

Jabar: Jurnal Pendidikan Matematika, vol. 12, no. 2, pp. 447–455, 2021.

[15] P. El Kafrawy, A. Mausad, and H. Esmail, “Experimental comparison of methods for multi-label classification in different application domains,” *International Journal of Computer Applications*, vol. 114, no. 19, pp. 1–9, 2015.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, and F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020.

[17] A. Mahajan, D. Shah, and G. Jafar, “Explainable AI approach towards toxic comment classification,” in *Emerging technologies in data mining and information security*. Springer, 2021, pp. 849–858.

[18] G. I. Pérez-Landa, O. Loyola-González, and M. A. Medina-Pérez, “An explainable artificial intelligence model for detecting xenophobic tweets,” *Applied Sciences*, vol. 11, no. 22, pp. 1–27, 2021.

[19] Ł. Górski and S. Ramakrishna, “Explainable artificial intelligence, lawyer’s perspective,” in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 2021, pp. 60–68.

[20] I. B. Muktyas, Sulistiawati, and S. Arifin, “Digital image encryption algorithm through unimodular matrix and logistic map using Python,” in *AIP Conference Proceedings*, vol. 2331, no. 1. AIP Publishing LLC, 2021, pp. 020 006–1–020 006–7.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should I trust you?” Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[22] F. Z. Tala, “A study of stemming effects on information retrieval in Bahasa Indonesia,” Master thesis, Universiteit van Amsterdam, 2003.