Cyberbullying identification on twitter using random forest classifier

To cite this article: N Novalita et al 2019 J. Phys.: Conf. Ser. 1192 012029

View the article online for updates and enhancements.
Cyberbullying identification on twitter using random forest classifier

N Novalita¹, A Herdiani², I Lukmana³, and D Puspandari⁴
School of Computing, Telkom University, Bandung, 40257, Indonesia
E-mail: ¹natasyanovalita@students.telkomuniversity.ac.id,
²anisaherdiani@telkomuniversity.ac.id,
³indraluk@telkomuniversity.ac.id,
⁴diyaspuspandari@telkomuniversity.ac.id

Abstract. Cyberbullying is a repeated act that is done to harass other people online in which it is common to happen amongst adolescents although there are some who don’t understand the harm of cyberbullying. The incomprehension of cyberbullying itself can cause ignorance of cyberbullying’s harmful effects. By using a classification technique, a cyberbullying identification can be done on social media to provide a better understanding regarding this particular activity. In Indonesia, researches that involved cyberbullying have not solved the problem related to identification of tweets containing cyberbullying and non-cyberbullying content. Therefore, this research proposed the usage of Random Forest’s classification to help in identifying cyberbullying on one of the affected social media, Twitter. The proposed method was applied on 50 group of tweets using training/test split. The usage of Random Forest in this research is optimized by the consideration of general rules from previous study with some needed adjustments. Random Forest results in the highest F1-Score which is 0.90. The wrong prediction is caused by the inconsistency in which some rules that determine whether a group of tweets are considered to be cyberbullying the most are justified as non-cyberbullying in some cases as well as the other way around.

1. Introduction
Bullying activity especially on social media (cyber-bullying) in Indonesia, has increased rapidly. Cyberbullying occurs because of the misuse of social media [1] such as Twitter where its use as a means of communication between people has led to an increase in cyberbullying activities, especially among adolescents [2]. In fact, the results of a research supported by UNICEF and the Ministry of Communication and Information found that 58% of 43,500 adolescents do not understand about the dangers of cyberbullying [3] because the meaning of cyberbullying itself is still vague. Therefore, it is necessary to identify cyberbullying on social media that can help in providing a better understanding to cyberbullying.

In Indonesia, cyberbullying’s researches on Twitter have only been done several times where the focus of those researches are mostly around the pattern of words of cyberbullying, and clustering [4,5]. Unfortunately, those researches have not solved the problem related to identification of tweets containing cyberbullying and non-cyberbullying content.

Cyberbullying in Twitter can be seen in a form of repetitive tweets that contain harsh or negative words that were posted to insult other people [6]. However, not all tweets that contain abusive words can be categorized as cyberbullying (Example: the word 'anjing' is usually used to insult someone but
the context will be different if the word ‘anjing’ is meant to be used to call an actual animal, etc.). Therefore, some rules that determine cyberbullying behaviour can also help in distinguishing some tweets containing cyberbullying and non-cyberbullying content [7] in which it can give different results in the application of cyberbullying.

Based on the problems that are stated above, this research proposed the usage of classification method in order to identify cyberbullying on Twitter. Sessionization will be carried out at an early stage to see the repetition of cyberbullying. Feature extraction will also be done by observing general rules from the previous study [7] with some needed adjustments that can help in identifying cyberbullying in accordance with the criteria on Twitter for data in Indonesian before applying Random Forest’s classification method in the cyberbullying identification system. The supervised learning method will be operated by building numerous decision trees at training time and come up with a classification needed to identify cyberbullying on the collected dataset. Random Forest had a high accuracy in performance based on the previous study which was using English tweets as the dataset [6].

2. Related works
Cyberbullying is every action which are done through electronic media or other digital media by an individual or groups that are harmful, abusive, and tend to be aggressive in order to disturb others [8]. In Indonesia, Cyberbullying is defined in the ITE Law as an action that is closely related to defamation [9]. Article 27 paragraph (3) of the ITE Law states that "Every person intentionally and without rights distributes and / or transmits and / or accesses Electronic Information and / or Electronic Documents that contains insults and / or defamation". Then in Article 27 paragraph (4) the ITE Law states that "Every person intentionally and without rights distributes and / or transmits and / or accesses Electronic Information and / or Electronic Documents that contains extortion and / or threats".

Identification of cyberbullying is formulated as a classification problem and involves document classification, topic detection, and sentiment analysis to detect bullying by looking at the characteristics of the message, the sender, and the recipient of the message [10].

Previously, there had been several studies related to cyberbullying which involved cyberbullying cases in foreign languages on several social media using various NLP and Text Mining techniques. Dinakar [11], Sanchez [12], Perez [13], and Dadvar [14] used various themes such as race, culture, and so on as a topic of general bullying to detect cyberbullying. The technique was implemented in the data collection stage for this final project. Despoina [6] detected cyberbullying and aggression on Twitter by paying attention to the main characteristics of cyberbullying, namely power differential and repetitiveness and supported by Random Forest where the precision and recall results were 89.9% and 91.7%.

In other studies, Margono [4,5] analysed Indonesian’s words of bullying from Twitter using the Association Rules and FP-Growth as well as clustering in both studies with similar topics.

Sarna [7] used machine learning to detect perpetrators of cyberbullying based on the chance of finding cyberbullying that is done directly or indirectly. They also compiled eight rules that can be used as features on the classification process of cyberbullying tweets.
3. Proposed method
In this section, the system design of the current research will be elaborated. The steps of the process are as shown in figure 1.

![Overall process](image)

Figure 1. Overall process.

3.1. Twitter dataset
The dataset is taken from Twitter using Twitter streaming API. Crawling data is carried out during January - March period of time in 2018.

Keywords that are used to assist in data collection relate heavily to words that are widely used for bullying such as “bangsat”, “anjing”, and so on. This method follows the step from previous research in which the Gender Bullying analysis was conducted based on LGBT-related keywords such as "gay" and "bitch" [12]. Then, to add variations in data, keywords like 'lgbt', 'persib', and so on are also used. This data is compiled in CSV file.

3.2. Sessionization
The grouping of a number of tweets at this stage is necessary in distinguishing whether the user behaves as a bully [6] based of cyberbullying’s repetitive aspect. For each session, the inter-arrival time between tweets does not exceed a predefined time threshold of 8 hours [6].

Each session will be divided into groups where each group has 5-10 tweets from the same user to analyze. Groups that have less than 5 tweets are eliminated.

3.3. Data Labeling
The next stage includes data labeling which is done by labelling each group of existing tweets as non-cyberbullying, or cyberbullying content. The following is a description of a group of non-cyberbullying and cyberbullying tweets’ criteria on Twitter based on previous research that can help participants in labeling session to determine the labels [8,15]:
a. Non-Cyberbullying. There are no tweets related to negative meaning at all or there are only some tweets that have harsh words or have negative meanings with a maximum of 2 tweets which is claimed as cyberaggressor.

b. Cyberbullying. There are various tweets or retweets (usually the number of tweets that have negative meanings are more than 2) on the same topic and are done repeatedly with the aim of insulting other users.

The examples of non-cyberbullying and cyberbullying group of tweets can be seen on table 1.

| Examples of Non-Cyberbullying | Examples of Cyberbullying |
|-------------------------------|---------------------------|
| User1 | UserA | Tweet | UserA | Tweet |
| Ada apa dengan kerusuhan Arema Persib? | @userb adu bangsat dah lu |
| RT @User2: Pelatih Persib Bandung, Mario Gomez, mengalami luka di kepalanya dalam kerusuhan pertandingan melawan Arema FC. | @userw wah bangsat lu emang |
| @User3 @User4 Bangst eta amankeun emot | @userb ah gimana si lu anjing |
| Jadi nanti gmn ya pertandingannya | @userb makanya kalo ngomong mikir, tolol |
| @User5 Semangat yaaa | @userb lu kaya anjing sih goblok |

8 rules that were obtained in previous studies [7] are also used as a criteria that can help participants to label the data as well as distinguishing cyberbullying and non-cyberbullying that can be applied in Indonesian language, i.e.:

1. The number of bad words in the tweet (e.g. 'anjing', 'pelacur', etc). The list of bad words are collected from previous studies [5] and from youswear.com [18].

2. The number of words that show negative feelings (e.g. 'bohong', 'benci', 'cacat', etc). The list of negative words that are used refers to the results of previous studies [19,20]. However, there are some non-negative words in the word list, so the word list is adjusted as needed.

3. The number of words that show positive feelings (e.g. 'bagus', 'senang', 'menakjubkan', and others to distinguish non-cyberbullying and cyberbullying tweets. The word list that are used refers to the results of previous studies [19,20] and is adapted to the existing data.

4. Combination of the first pronoun with words that show negative feelings and second person pronouns (e.g. 'Saya tidak suka kamu').

5. Combination of the second pronoun with a bad word (e.g. 'Lo itu pelacur').

6. Combination of the first pronoun, negative words and third pronouns or names (e.g. 'Saya tidak percaya mereka'). Username is also taken into consideration in this rule because the involvement of mentioning the username (or can be referred to as 'mentions' on Twitter) is considered important and can affect the results of feature extraction. This consideration is strengthened by the results of several testings which can be seen in A2 and A3 on Appendix’s page.

7. Combination of third pronouns with bad words or names (e.g. 'Mawar itu bangsat!'). Same with the previous rule, username is also taken into consideration in this rule. This consideration is strengthened by the results of several testings which can be seen in A2 and A3 on Appendix’s page.

8. A combination of URLs, bad words and pronouns (e.g. 'Kamu memang anjing https://t.co/m3h5xj00'). Same with the previous rule, username is also taken into consideration in
this rule. This consideration is strengthened by the results of the tests which can be seen in A2 and A3 on Appendix’s page.

These rules are also used in section 3.4, 'Feature Extraction'. Participants do manual labeling using Google Sheet. Because of the different levels of confidence between each participant regarding determining the existing tweets as cyberbullying or the other way around, the system will provide 4 choices based on Likert measurements to eliminate 'neutral' choices [16] as shown in table 1. The four choices are divided into two groups to get the 'cyberbullying' and 'non-cyberbullying' labels. Options 1 and 2 are used to measure the number of 'non-cyberbullying' scores, while options 3 and 4 are used to measure the number of 'cyberbullying' scores.

| Options(i) | Point | Description          |
|-----------|-------|----------------------|
| 1         | -2    | Very Non-Cyberbullying|
| 2         | -1    | Non-Cyberbullying     |
| 3         | 1     | Cyberbullying         |
| 4         | 2     | Very Cyberbullying    |

Table 2. Labeling options details.

Each group will be labelled by a minimum of three participants to overcome the value bias. The stages to get the final labels for each group are described as follows [17]:

1. As shown in table 2, points for option 1 is greater than option 2 to represent the level of non-cyberbullying that is higher than option 2. The minus sign is used to distinguish between the 'cyberbullying' and 'non-cyberbullying' groups.

2. Measurement of the score of the 'cyberbullying' label (C) is calculated using equation (1). Option 3 or 4 will be multiplied by the number of participants who choose that option.

\[
C = \frac{\sum_{i=3}^{4} point_i \times total\_participants}{\sum_{i=1}^{4} point_i \times total\_participants_i}
\]

3. Calculate the odds of the 'non-cyberbullying' label (NC) using equation (2).

\[
NC = 1 - C
\]

The result of sessionization and data labeling can be seen on A1 on Appendix’s page.

### 3.4. Preprocessing

Preprocessing is done to eliminate enough noises and get the desired results before it can be processed further. The preprocessing stages of the data are as follows:

1. Data Cleaning
   - At this stage RT, symbols, and emoticons are deleted to help tidying up the data.

2. Case Folding
   - This stage is used to convert each existing uppercases into lowercases.

An example of preprocessing’s result can be seen in table 3.

| Table 3. Example of preprocessing’s result. |
|--------------------------------------------|
| Sentence | RT @userc: WAH Kamu dan rekan2 kerjamu disitu sudah seperti bangsat buat mempermasalahkan hal yang sepele! @user |
| Data Cleaning and Case Folding’s result | userc : wah kamu dan rekan2 kerjamu disitu sudah terlalu bangsat buat mempermasalahkan hal yang sepele @user |
3.5. Features extraction
When the preprocessing step is finished, it is necessary to extract features to determine the features of the set of words (tokens) as an input for the classification process where feature extraction itself will be divided based on 8 aspects that can be used as rules based on previous research [7].

Each detected rule will be calculated where the system will add up the number of occurrences of each aspect in the existing tweets. An example of the application of feature extraction can be seen in table 4.

| No | Tweet | User name | Rule 1 | Rule 2 | Rule 3 | Rule 4 | Rule 5 | Rule 6 | Rule 7 | Rule 8 | Label |
|----|-------|-----------|--------|--------|--------|--------|--------|--------|--------|--------|-------|
| 1  | @user1 Anjing kau | User3 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2  | Bangsat, tolol dia. Gue gak suka sama dia. | User3 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 3  | Besok bakal terbit komentar yg Nyiyirin Presiden https://t.co/m3h5 | User5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 4  | Saya bingung ada elite politik tp dungu, @jokowi tidak butuh pengakuan dr Gerindra | User5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

At the example above there are several tweets used in the feature extraction system. The label '0' is the 'Cyberbullying' label and the label '1' is the 'Non_Cyberbullying' label.

After getting the needed results as in the example above, the sum of each rules for each group of tweets is made based on the username where the tweets written by user3 go to 'BATCH_01' while the tweets written by user5 go to 'BATCH_02'. The end result of this feature extraction will be used as an input in the Random Forest classification system. In this research, 50 groups of tweets are collected. Examples of feature extraction per group can be seen in table 5.

Table 5. Example of features extraction result per group.

| Group of Tweets | Rule 1 | Rule 2 | Rule 3 | Rule 4 | Rule 5 | Rule 6 | Rule 7 | Rule 8 | Label |
|----------------|-------|--------|--------|--------|--------|--------|--------|--------|-------|
| BATCH_01       | 3     | 0      | 0      | 0      | 1      | 0      | 1      | 0      | 0     |
| BATCH_02       | 1     | 1      | 0      | 0      | 0      | 0      | 0      | 1      | 0     |

3.6. Random Forest Classification
Random Forest Classification is done to get the final result of the identification system. A module that provides learning algorithms and is named Scikit-learn [21] in Python programming is used at this stage to process the algorithm. Feature extraction’s results are inputs used in this classification.
Here are the following steps in building the Random Forest classification model:

1. Distribution of datasets into training data and data testing is done by splitting the data to get 80% for data training and 20% for data testing while using random_state = 10. The random_state variable is determined to get the same results every time the system is run.

2. Random Forest Classification is done by using 17, 20, 50, 57, 93, and 100 trees while using random_state = 9. The random_state variable is determined to get the same results every time the system is run.

3. Determine max_features to find out the number of features considered to get the best split. The max_features value matches the overall value of the attributes in the used dataset. In this final project, max_features is set by default according to the Scikit Learn module where max_features = "auto" and max_features = sqrt (n_features) is calculated. Random Forest will randomly select features from the total max_features (M) where k < M before those features are used to construct a decision tree.

4. The construction of the decision tree in the Random Forest is assisted by gini criterion to select the best attribute as root node. The construction of decision trees involves data training and features.

5. A prediction will be obtained by observing the majority voting of each prediction class on the tree that has been arranged in the system. The system prediction results are then compared with the results of predictions that have been made on the previous data labeling for evaluation. However, in the Scikit Learn module, the Random Forest classification model tends to do soft voting where the prediction results are obtained from prediction probability calculations.

The following is an example of Random Forest classification model’s usage by utilizing 15 datasets by using 80% for data training and 20% for data testing and the number of trees that are built is 5 trees. Classification Model Random Forest selects 2 rules (k) from the total 8 rules (M) to create one of the decision trees that can be seen in figure 2.

![Figure 2. Example of one decision tree.](image-url)
Table 6. Example of predictions made by each decision trees.

| Group of Tweets | Tree 1 | Tree 2 | Tree 3 | Tree 4 | Tree 5 |
|-----------------|-------|--------|--------|--------|--------|
| BATCH_03        | 1     | 0      | 1      | 1      | 1      |
| BATCH_08        | 0     | 0      | 1      | 1      | 0      |
| BATCH_13        | 0     | 1      | 1      | 0      | 0      |

Scikit Learn’s Random Forest uses a soft majority voting where the prediction with the highest prediction probability are chosen as the result of the final predictions by the system. This soft majority voting is chosen because it is more accurate than majority voting in general [22]. Prediction probability is useful for determining the final prediction results and is calculated by dividing a prediction class with the overall prediction results in the entire trees as shown in equation (3).

\[
\text{Prediction Probability} = \frac{\text{The number of a specific prediction (class)}}{\text{The number of trees}}
\]  

The class in question is the 'Cyberbullying' or 'Non_Cyberbullying' labels themselves.

Table 7. Example of predictions made by each decision trees.

| Group of Tweets | Prediction Probability | Predictions |
|-----------------|------------------------|-------------|
|                 | Cyberbullying | Non_Cyberbullying |             |
| BATCH_03        | 0.20       | 0.80            | Non_Cyberbullying |
| BATCH_08        | 0.60       | 0.40            | Cyberbullying |
| BATCH_13        | 0.60       | 0.40            | Cyberbullying |

From table 7 it can be seen that the highest prediction probability in BATCH_03 is on the 'Non_Cyberbully' label so the final prediction obtained is 'Non_Cyberbullying'. As for BATCH_08 and BATCH_13, the highest prediction probability is found on 'Cyberbullying' label so that the final prediction results obtained are 'Cyberbullying'.

4. Result and Analysis
Accuracy calculation is needed to determine whether the classification model has performed well or the other way around where the parameters needed to measure the performance itself consist of precision, recall, and f1-score. In its use 4 terms are known, namely [23]:

1. True Positive (TP). Classified data has a number of values that are predicted to be positive and the fact is also holding a positive meaning.

2. True Negative (TN). Classified data has a number of values that are predicted to be negative and they are actually negative.

3. False Positive (FP). Classified data has a number of values that are predicted to be positive but they are actually negative.

4. False Negative (FN). Classified data has a number of values that are predicted to be negative and the fact is supposed to be positive.

Precision, Recall, and F1-Score can be calculated using these following ways [24]:
1. Precision is the percentage of correctly predicted documents and they are proved as the truth as well. Precision is used to measure the exactness of the predictions made by the system. The calculation of precision can be done by following equation (4).

\[
Precision = \frac{TP}{TP + FP}
\]  

2. Recall is the percentage of correct documents and the right predictions are obtained as well. Recall is used to measure the item’s quantity that is correct and it is able to be predicted correctly. The calculation of recall can be done by following equation (5).

\[
Recall = \frac{TP}{TP + FN}
\]  

3. F1-Score is used to get the weighted average between precision and recall. The calculation of F1-Score can be done by following equation (6).

\[
F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision}
\]  

The predicted results that have been determined by Random Forest classification model with a number of different trees provide quite a variety of results. The performance calculation of the Random Forest classification model can be seen on figure 3.

![Figure 3. The result of random forest’s performance.](image)

Based on figure 3, the highest F1-score obtained is 0.90 in the classification model that uses 50, 57, 93, and 100 trees. This proves that the use of the number of trees can affect the results of performance and the use of odd number of trees with even number data (50 datas) cannot ensure that the results will increase because the use of an even number of trees with even number of data also produces similar results. The difference in the prediction results of Random Forest classification on 17 number of trees with F1-Score 80% and the prediction results on the number of trees as much as 50 with the highest F1-Score 90% can be seen in table 8 and table 9.
To ensure this, another test was carried out with a number of different trees which obtained F1-Score but the actual label given is 'Non_Cyberbullying' and vice versa. Some rules useful for distinguishing cyberbullying and non-cyberbullying actions based on the detection of people's pronouns combined with harsh words and negative words are seen in A4, A5, and A7. The content of BATCH_38 can be seen in A4.

From tables 8 and 9, a prediction error was found on BATCH_38 where the actual result shows the 'Cyberbullying' label but the Random Forest classification model gives a prediction of 'Non_Cyberbullying'. The content of BATCH_38 can be seen in A4 on Appendix's page. A prediction error was found in BATCH_43 which shows the actual result of 'Non_Cyberbullying' but the Random Forest classification model gives a prediction of 'Cyberbullying'. The content of BATCH_21 is predicted as 'Non_Cyberbullying' where only Rule 2 (number of negative words) is detected from the batch. While the results of the BATCH_23 test are predicted to be 'Cyberbullying' even though only Rule 2 and Rule 3 (number of positive words) are detected from the batch. This proves that some rules that represent cyberbullying actions on a tweets can necessarily be predicted as 'Cyberbullying' and vice versa because of the inconsistencies described earlier where this can be caused by a number of groups of tweets that only the respondents can perceive as defamatory without noticing the rules detected but the classification model fails to understand because the model only sees the detected pattern of rules. Examples of tweets that get wrong predictions from the rules can be seen in A4, A5, and A7 on Appendix's page.

Table 8. Random forest classification’s predicted result with 17 trees on 20% of data testing.

| Batch   | Rule | Actual Label  | Predicted Label |
|---------|------|---------------|-----------------|
| BATCH_38 | 4 3 1 0 0 1 2 1 | Cyberbullying | Non_Cyberbullying |
| BATCH_24 | 11 0 0 4 0 6 0 | Cyberbullying | Cyberbullying |
| BATCH_45 | 0 0 0 0 0 0 0 0 | Non_Cyberbullying | Non_Cyberbullying |
| BATCH_43 | 2 1 2 0 0 0 2 1 | Non_Cyberbullying | Cyberbullying |
| BATCH_48 | 0 0 0 0 0 0 0 0 | Non_Cyberbullying | Non_Cyberbullying |
| BATCH_21 | 0 2 2 0 0 0 5 0 | Cyberbullying | Cyberbullying |
| BATCH_04 | 0 3 3 0 0 0 0 0 | Non_Cyberbullying | Non_Cyberbullying |
| BATCH_31 | 5 2 0 0 0 0 5 0 | Cyberbullying | Cyberbullying |
| BATCH_08 | 3 5 0 0 0 0 3 1 | Cyberbullying | Cyberbullying |
| BATCH_07 | 5 0 0 0 0 0 5 0 | Cyberbullying | Cyberbullying |

Table 9. Random forest classification’s predicted result with 50 trees on 20% of data testing.

| Batch   | Rule | Actual Label  | Predicted Label |
|---------|------|---------------|-----------------|
| BATCH_38 | 4 3 1 0 0 1 2 1 | Cyberbullying | Non_Cyberbullying |
| BATCH_24 | 11 0 0 4 0 6 0 | Cyberbullying | Cyberbullying |
| BATCH_45 | 0 0 0 0 0 0 0 0 | Non_Cyberbullying | Non_Cyberbullying |
| BATCH_43 | 2 1 2 0 0 0 2 1 | Non_Cyberbullying | Non_Cyberbullying |
| BATCH_48 | 0 0 0 0 0 0 0 0 | Non_Cyberbullying | Non_Cyberbullying |
| BATCH_21 | 0 2 2 0 0 0 5 0 | Cyberbullying | Cyberbullying |
| BATCH_04 | 0 3 3 0 0 0 0 0 | Non_Cyberbullying | Non_Cyberbullying |
| BATCH_31 | 5 2 0 0 0 0 5 0 | Cyberbullying | Cyberbullying |
| BATCH_08 | 3 5 0 0 0 0 3 1 | Cyberbullying | Cyberbullying |
| BATCH_07 | 5 0 0 0 0 0 5 0 | Cyberbullying | Cyberbullying |

From tables 8 and 9, a prediction error was found on BATCH_38 where the actual result shows the 'Cyberbullying' label but the Random Forest classification model gives a prediction of 'Non_Cyberbullying'. The content of BATCH_38 can be seen in A4 on Appendix's page. A prediction error was found in BATCH_43 which shows the actual result of 'Non_Cyberbullying' but the Random Forest classification model provides a prediction of 'Cyberbullying'. The content of BATCH_43 can be seen in A5 on Appendix's page.

This error is derived from the inconsistency of detected rules’ patterns for each category. The inconsistency itself can be seen from several groups of tweets where Rule 4 until Rule 7—in which they are some rules useful for distinguishing cyberbullying and non-cyberbullying actions based on the detection of people's pronouns combined with harsh words and negative words—are sometimes detected but the actual label given is 'Non_Cyberbullying' and vice versa.

To ensure this, another test was carried out with a number of different trees which obtained F1-Score 60%. Test results can be seen in A6 on Appendix’s page.

From the result it can be seen that BATCH_27 is predicted as 'Non_Cyberbullying' where only Rule 2 (number of negative words) is detected from the batch. While the results of the BATCH_23 test are predicted to be 'Cyberbullying' even though only Rule 2 and Rule 3 (number of positive words) are detected from the batch. This proves that some rules that represent cyberbullying actions on a tweets can necessarily be predicted as 'Cyberbullying' and vice versa because of the inconsistencies described earlier where this can be caused by a number of groups of tweets that only the respondents can perceive as defamatory without noticing the rules detected but the classification model fails to understand because the model only sees the detected pattern of rules. Examples of tweets that get wrong predictions from the rules can be seen in A4, A5, and A7 on Appendix’s page.
The overall test results can be seen in A2 and A3 on Appendix’s page. This research has shown that Random Forest is able to give a high result which is better compared to the previous study [17] given the fact that Random Forest does better in accuracy because of their excellent performance on low dimensional datasets such as stated in another study [24] and the adjustments on general rules such as the consideration of username also gave a better performance as seen in A2 and A3 on Appendix’s page.

The inconsistency that is caused by the respondents’ understanding without even following the rules at hand yet the classification model fails to understand proved that general rules which considers bad words, negative words, and the use of pronouns aren’t enough to distinguish between cyberbullying and non-cyberbullying content given the fact that human-based logic hasn’t been injected into it. Other features may be needed in order to avoid such inconsistency in the future.

5. Conclusion

Based on the completion of the final project that has been done, it is known how to identify both cyberbullying and non-cyberbullying tweets. Judging from the testing’s results with the Random Forest classification, it is found that the system identified cyberbullying tweets successfully with the best F1-Score of 0.90.

A 10% prediction error was obtained from the pattern of detection of rules from tweets about cyberbullying or non-cyberbullying that is inconsistent where Rule 4 - Rule 7 in which they are the rules useful for distinguishing cyberbullying and non-cyberbullying actions based on the detection of combinations of pronouns with harsh words and negative words are detected from the group of tweets a few times but the actual label given is 'Non_Cyberbullying' label and vice versa. This can be caused by the group of tweets storing meanings that can be understood by respondents as something that pollutes the good name or vice versa without paying attention to the rules that are detected but the classification model fails to understand because the model only sees the detected pattern rules.

The following suggestions can be made for further research:

1. Implementing term-weighting in each rules so that the detected rules can have a calculated weight and make it easier for the system to distinguish tweets that are categorized as cyberbullying and non-cyberbullying to avoid prediction errors and produce more accurate performance.

2. The usage of sarcasm (a bitter remark that may employ ambivalence) that is directed to insult people in which they are not included in the general rules should be considered in order get more in depth understanding about cyberbullying and its identification on social media.

3. Power differentials (one of the main factor that differentiate cyberbullying content with non-cyberbullying content that includes anonymity, the constant possibility of threat, and a potentially larger audiences), user (age, gender, sexual orientation, race etc.), and network features (number of friends, uploads, likes, etc.) should be taken into consideration to get more in depth understanding about cyberbullying and its identification on social media.

4. Adding a system for future research that can interact with users of cyberbullying itself to be socialized about the dangers of cyberbullying and provides support to help the victims of cyberbullying.
Appendix

A1. Sessionization and Data Labeling’s Result

The overall result of sessionization and data labeling can be seen by opening this link: [http://tiny.cc/r79fwy](http://tiny.cc/r79fwy)
A2. Performance’s Calculation I
Using Train/Test Split

| Data Partition | 70:30 | 80:20 | 90:10 |
|----------------|-------|-------|-------|
| Tree           | 17    | 20    | 50    |
|                | 57    | 93    | 100   |
| F1-Score       | 66%   | 66%   | 66%   |
| Username is Considered On Rules |
| F1-Score       | 72%   | 79%   | 79%   |
| Username is Not Considered On Rules |

A3. Performance’s Calculation II

| k-Fold | 5 | 10 | 15 | 20 |
|--------|---|----|----|----|
| Tree   | 17 | 20 | 50 | 93 | 100 |
| F1-Score | 74% | 78% | 76% | 76% | 76% |
| Username is Considered On Rules |
| F1-Score | 68% | 66% | 64% | 61% | 64% |
| Username is Not Considered On Rules |

Using Cross Validation

A4. BATCH_38’s Contents

| Tweet | Rule | Actual Label | Predicted Label |
|-------|------|--------------|-----------------|
| @DIEsPLAYGROUND @SOOJUNGdIEs @RYUWONDIEs BANGSAT KALIAN RIBUTRIBUT TIDAK PENTING https://t.co/JRIH7wphTO | 1: (bangsat) 1: {tidak pening} 0 0 0 | 1: {@user - bangsat} | 1: {@user - bangsat - link url} |
### A5. BATCH_43's Contents

| Tweet | Rule | Actual Label | Predicted Label |
|-------|------|--------------|-----------------|
| @Tickno4 Biarkan anjing manggonjong Jokowi tetap melenggang https://t.co/Be9IH71LQu | 1: {anjing} 2: {setuju, pantas} | Cyberbully | Non_Cyberbully |
| @Tickno4 Tetap jokowi https://t.co/Z7rKZ8OQqC | 1: {Jokowi-anjing} 1: {Jokowi-anjing-link} | Cyberbully | Non_Cyberbully |
| @PribumiZindaHai Jokowi Saya tidak pantas jadi Capres karena jadi Yg SETUJU jadiin Jokowi ketua di kampungnya Balikin ke Solo | 1: {saya-setuju} | Cyberbully | Non_Cyberbully |
| RT @RiskaAmeliaG 30: Proyek Strategis Presiden Jokowi senilai Rp 948 Triliun telah Rampung | 1: {saya-setuju} | Cyberbully | Non_Cyberbully |
war gitu tapi emang goblok aja di rl nya wkwkwk

### A6. The Result of Random Forest Classification Using 17 Trees with Different random_state.

| Batch      | Rule | Actual Label       | Predicted Label |
|------------|------|-------------------|-----------------|
| BATCH_14   | 1    | Non_Cyberbullying | Cyberbullying   |
| BATCH_40   | 1    | Non_Cyberbullying | Cyberbullying   |
| BATCH_31   | 1    | Cyberbullying     | Cyberbullying   |
| BATCH_46   | 1    | Cyberbullying     | Cyberbullying   |
| BATCH_18   | 1    | Non_Cyberbullying | Non_Cyberbullying |
| BATCH_49   | 1    | Cyberbullying     | Cyberbullying   |
| BATCH_27   | 1    | Cyberbullying     | Non_Cyberbullying |
| BATCH_26   | 1    | Cyberbullying     | Cyberbullying   |
| BATCH_23   | 1    | Non_Cyberbullying | Cyberbullying   |
| BATCH_24   | 1    | Cyberbullying     | Cyberbullying   |

### A7. BATCH_27’s Contents

| Tweets | Rules | Actual Label       | Predicted Label |
|--------|-------|-------------------|-----------------|
| @GOALID SANKSI MACAM APA NI PSSI LEMBEK KALI MACAM LONTONG KEMAREN SORE PERSI PERSIJA PSM SAJA KAU GEMBOSI https://t.co/sZCqAg5avg | 0 | Cyberbullying     | Non_Cyberbullying |
| @detikcom SANKSI MACAM APA NI PSSI LEMBEK KALI MACAM LONTONG | 0 | Cyberbullying     | Non_Cyberbullying |
KEMAREN SORE
PERSI PERSIJA PSM
SAJA KAU GEMBOSI
https://t.co/zGQdE8Ssks
@BolaNet
SANKSI
MACAM APA NI PSSI
LEMBEK KALI
MACAM LONTONG
KEMAREN SORE
PERSI PERSIJA PSM
SAJA KAU GEMBOSI
https://t.co/zGQdE8Ssks
@detiksport
SANKSI
MACAM APA NI PSSI
LEMBEK KALI
MACAM LONTONG
KEMAREN SORE
PERSI PERSIJA PSM
SAJA KAU GEMBOSI
https://t.co/3dZgmj9ar0
@vfrontlinepc
@officialvpc
persi
References

[1] Hidajat M, Adam A R, Danaparamita M and Suhendrik 2015 Dampak media sosial dalam cyberbullying 2015 *ComTech: Computer, Mathematics, and Engineering Applications* 6 no 1 pp 72-81

[2] Sophia A 2015 Analysis of tweets related to cyberbullying: exploring information diffusion and advice available for cyberbullying victims 2015 *Int. J. of Cyber Behavior, Psychology and Learning* 5 no 4 pp 31-52

[3] Indonesia MCIT 2016 Digital citizenship safety among children and adolescents indonesian KOMINFO

[4] Margono H, Yi X and Raikundalia G K 2014 Mining indonesian cyber bullying patterns in social networks *Proc. of the Thirty-Seventh Australasian Computer Science Conf.* pp 115-124

[5] Margono H, Yi X and Raikundalia G 2015 Clustering indonesian cyberbullying words in social networks *Int. J. of Computer Science and Electronics Engineering* 3 no 4 pp 274-278

[6] Chatzakouy D, Kourtellisz N J B, De Cristofaro E, Stringhini G and Vakali A 2017 mean birds: detecting aggression and bullying on twitter 2017

[7] Sarna G and Bhatia M 2015 Content based approach to find the credibility of user in social networks: an application of cyberbullying 2015 *In Int. J. Mach. Learn. & Cyber*

[8] Tokunaga R S 2010 Following you home from school: A critical review and synthesis of research on cyberbullying victimization 2010 *Computers in Human Behavior* 26 pp 277 - 287

[9] Indonesian Government 2016 Undang-Undang Informasi dan Transaksi Elektronik

[10] Salawu S, He Y and Lumsden J 2017 Approaches to automated detection of cyberbullying: a survey 2017 IEEE

[11] Dinakar K, Reichart R and Lieberman H 2011 Modeling the detection of textual Cyberbullying the social mobile web 2011

[12] Sanchez H and Kumar S 2011 Twitter bullying detection 2011

[13] Pérez P J C, Valdez C J L, Ortiz M D G C, Barrera J P S and Pérez P F MISAAC: instant messaging tool for cyberbullying detection

[14] Dadvar M and De Jong F 2012 Cyberbullying detection: a step toward a safer internet yard 2012 *Int. Conf. Companion on World Wide Web*

[15] Grigg D W 2012 Cyber-aggression: definition and concept of cyberbullying 2012 *Australian J. of Guidance and Counselling* 20 no 02

[16] Allen I E and Seaman C A 2007 Likert scale and data analyses

[17] Nurrahmi H 2018 Indonesian twitter cyberbullying detection using text classification and user credibility 2018 *Int. Conf. on Information and Communications Technology (ICOACT)* pp 543-548

[18] "YourSwear.com : Swearing," [Online]. Available: http://www.youswear.com.

[19] Wahid D H and Azhar S N 2016 Peringkasan sentimen ekstrakatif di twitter menggunakan hybrid TF-IDF dan cosine similarity 2016 *Indonesian Journal of Computing and Cybernetics Systems* 10 no 2 pp 207-218

[20] Bing L, Hu M and Cheng J 2005 Opinion observer: analyzing and comparing opinions on the web 2005 *14th Int. World Wide Web Conference* Chiba

[21] Padregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D and Brucher M 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier Sci-Kit Learn [Online] Available: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[22] Breiman L 1994 Bagging Predictors Department of Statistics University of California Berkeley
[23] Vaitheeswaran G and Arockiam L 2016 Combining lexicon and machine learning method to enhance the accuracy of sentiment analysis on big data *Int. Journal of Computer Science and Information Technologies* 7 pp 306-311

[24] Caruana R, Karampatziakis N and Yessenalina A 2008 An empirical evaluation of supervised learning in high dimensions *The 25th Int. Conf. on Machine Learning*