Improved Model for Identifying the Cyberbullying Based on Tweets of Twitter

Darwin Samalo, Rizky Martin, Ditdit Nugeraha Utama
Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University
Jakarta, Indonesia
E-mail: darwin.samalo@binus.ac.id, rizky.martin@binus.ac.id, ditdit.utama@binus.edu

Keywords: cyberbullying, Twitter, natural language processing, decision tree, text mining

Received: November 23, 2022

The surge of cyberbullying on social media platforms is a major concern in today's digital age, with its prevalence escalating alongside advancements in technology. Thus, devising methods to detect and eliminate cyberbullying has become a crucial task. This research meticulously presents a refined model for identifying instances of cyberbullying, building on previous methodologies. The process of devising the model involved a thorough literature review, object-oriented design, and decision tree methodologies to shape the labelling procedure and build the classifier. Data pre-processing was executed using RapidMiner, considering six intrinsic components. The final model successfully classified Indonesian-language tweets into five distinct categories: animal, psychology and stupidity, disabled person, attitude, and general bullying, with an accuracy rate of 99.56%.

Povzetek: Predstavljen je izboljšan model za odkrivanje spletnega nadlegovanja na družbenih omrežjih. S pomočjo pregleda literature, objektno usmerjenega oblikovanja in odločitvenih dreves, model klasificira tvite v indonezijskem jeziku v pet kategorij z natančnostjo 99,56%, kar predstavlja pomemben korak v boju proti spletnemu nadlegovanju.

1 Introduction

The internet has become a part of life for most Indonesians. Datareportal states that Indonesia will have a population of 274.9 million population in January 2021. Of the total population, 202.6 million internet users are in Indonesia [1]. With the internet, we can search for information, work, shop and interact with other internet users, which can be done using social media. Social media is a tool that allows people to share information in the form of images, videos, and texts through specific networks [2]. Indonesia is one of the countries with the most significant number of social media users globally. Datareportal states that 170 million social media users will be in Indonesia in January 2021 [1]. That means 61.8% of the Indonesian population uses social media.

Various kinds of social media platforms are available in Indonesia, such as Twitter, Instagram, TikTok, and others. With more social media platforms nowadays, more crimes can be committed: cyberbullying. Like traditional bullying, cyberbullying is an adverse action intended to cause harm to a particular person or group by sending text messages, images, and videos to someone’s social media account to insinuate, insult, harass, and discriminate [3][4]. Many researchers have conducted research on cyberbullying based on social media, but these are only limited to classifying whether it is bullying or not. Most research does not classify what cyberbullying acts are bullies do. In this study, the researchers improve the algorithm proposed by the previous research conducted by Ade and Utama [5] to avoid potential unfair decision. The research classifies bullying into five: animal, psychology, stupidity, disabled person, attitude, and general.

Researchers use Twitter social media by collecting user posts, which will then be used for analysis to classify cyberbullying actions. Twitter post data crawling is done with RapidMiner. Mat, Lajis, and Nasir [6] said RapidMiner could create models more conveniently and quickly with their graphical visual environment feature. RapidMiner is open-source software (OSS) [7]. This study classifies cyberbullying by using the decision tree method. The decision tree is a classification method that uses a tree structure [8], where each node represents an attribute, its branch represents the attribute's value, and the leaves represent the class [9]. The researchers use the bag-of-words (BOW) model to extract the features used in the decision tree. BOW is one of the most popular text representation models [10]. BOW is often used to represent the semantics of text, where each word will be measured based on its weight [11]. This research aims to improve the algorithm for collecting and labelling Indonesian datasets and then classifying the types of cyberbullying using the decision tree method.

2 Related works

In recent years, many studies have been related to analysing and detecting cyberbullying using text mining by classifying posts. Haidar, Chamoun, and Serhrouchni [12] use Naïve Bayes and Support Vector Machine (SVM) to detect cyberbullying in the Arabic Language. The dataset is taken from Facebook and Twitter. The average score for Naïve Bayes is 90.5%, and SVM is 92.7%. This experiment classifies text into bullying or non-bullying class.

Isa [13] also conducted a machine-learning approach. They are using Naïve Bayes and SVM. The experiment has been carried out, and the result is that SVM with poly kernel has the highest accuracy than other methods, with 97.11% accuracy. In their research, they used the Formspring dataset from Kaggle. There is much pre-processing performed in this research. They perform data cleaning, tokenization, transform to lowercase, stop word removal, and stemming. This research also classifies data into two main classes: cyberbully and non-cyberbully. Regarding data balancing on the classification of 2 classes, they divide them into four classes (non-cyberbully, cyberbully level severity low, cyberbully level severity middle, cyberbully level severity high) and 11 classes (non-cyberbully, cyberbully level severity 1-10). However, still, this classification is only classified into two main classes.

Another research conducted by Nurrahmi and Nurjanah [14] also classifies tweets into two classes that are cyberbullying and non-cyberbullying. They perform data cleaning, tokenization, and part-of-speech (POS) tagging in a pre-processing step. Tools that they used for POS tagging is the Indonesian Language POS Tagger. They used a Twitter dataset taken from January to February 2017. The method that they used is K-nearest Neighbor (KNN) and SVM. The SVM method also outperforms KNN, with 67% accuracy.

In previous works, Febriany and Utama [5] identify cyberbullying posts in five classes. In this research, they proposed a new algorithm for classifying posts into five classes of cyberbullying (animal, psychology, and stupidity, disabled person, attitude, and general). This research also evaluates this algorithm. The evaluation uses machine learning, Naïve Bayes. The result of evaluation using Naïve Bayes is 99.15%.

Table 1 shows a comprehensive summary of the existing research on cyberbullying detection. Based on previous research studies, many researchers only identify cyberbullying in two classes, bullying or not. However, in the last paper, the research (ibid.) conducted by Febriany and Utama [5] developed a model to identify cyberbullying in five classes. The machine learning method used was Naïve Bayes. In this study, researchers improve the algorithm in the labelling process and use the decision tree method because the decision tree performs better on large datasets.

This study presents a novel approach in comparison to the previous research conducted by Febriany and Utama [5], as it introduces an additional variant feature. In the previous study, if there were multiple maximum values for a label, the model selected the label in first place, whereas in this study, the label with the most variant words is chosen. The addition of the variant feature reduces the likelihood of unfair decisions that may have occurred in the previous study, where only the maximum value was considered. This is a significant contribution as it ensures fair decisions in cyberbullying classification. Furthermore, this study employs a decision tree as a classification method, which is advantageous for handling large datasets.

Table 1: Related works

Author	Tittle	Method	Result
Haidar, Chamoun, and Serhrouchni [12]	Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content	Machine Learning (ML) and Natural Language Processing (NLP)	Naïve Bayes is 90.5%, and SVM is 92.7%
Isa, Ashianti, et al. [12]	Cyberbullying Classification using Text Mining	Naïve Bayes and SVM	Naive Bayes is 92.81%, and SVM is 97.11%
Nurrahmi and Nurjanah [14]	Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility	KNN and SVM	The SVM method also outperforms KNN, with 67% accuracy
Febriany and Utama [5]	Analysis Model for Identifying Negative Posts Based on social media	Naïve Beyes	The result of using Naïve Beyes is 99.15%

3 Research methodology

The research was performed in six main stages: literature studying, data collecting and pre-processing, manual labelling, feature extracting, and classifying. Figure 2 explains the general description of processing data.

First, the label was obtained from research conducted by Ade and Utama (ibid.). After the labels are defined, the data is crawled from Twitter using RapidMiner based on the labels. The data that has been collected is continued to the pre-processing process for cleaning. The pre-processing process is done twice, the first while crawling data using RapidMiner, and the second using python. After the data is clean, the data will be labelled. However, to make machine learning make the classification easier, the data will be converted to a number. That can be done with feature extraction using Bag of Words (BoW) and Term Frequency Inverse Document Frequency (TF-IDF). After extracting the features, the classification will be done using a decision tree.

3.1 Literature study

The first step conducted in this research is a literature review study. Literature studies are carried out to deepen the author’s knowledge so that researchers can improve the proposed algorithm. The label used was also taken from the study literature, which the previous study also used this label.

3.2 Data collecting and pro-processing

The data used in this study were tweets posted by Twitter users in Indonesia. The Twitter data were crawled by operating the tool RapidMiner. The collected Twitter posts are limited by the labels from research conducted by Ade and Utama (ibid.). Ten thousand seven hundred eighty-six data have been obtained from data crawling. After the data were academically collected, the data were later pre-processed. Pre-processing aims to get clean and ready data to use in the model. Several steps were taken when conducting the data collecting and pre-processing (with scheme configured in Figure 1): retrieving Twitter, searching Twitter, selecting attributes, removing duplicates, filtering examples, sub-process, and writing CSV.

In the first step, RapidMiner is linked to a Twitter account to retrieve Twitter users’ posts. After the Twitter account is connected to RapidMiner, the search Twitter operator is used to find all tweets that contain a specific phrase. This study's collected data was based on the labels collected from a previous study (ibid.).

In the result of crawling data, there are several attributes such as 'text', 'from-user-id', 'created-at', and 'language'. Only the 'text' attribute is used for this research, and the unneeded attributes will be removed. The 'text' attribute contains the tweets of a Twitter user containing the label word that will be used. Afterwards, duplicates based on the 'text' attribute are removed using the 'remove duplicate' operator.

The missing value in the ’text’ attribute will be removed using the filter example’s operator. After removing the missing value, the string ’RT’ contained in the ’text’ attribute will also be removed using replace operator in the subprocess operator. After that, store the results in the ’CSV’ extension, which will proceed to the pre-processing stage using python.

After the pre-processed in RapidMiner, the data will also be pre-processed using Python. Pre-processing includes cleaning data, case folding, stop word removal, stemming, and tokenization.

The first pre-processing is to clean the data. The purpose is to delete data with no value and words without meaning, such as retweets, hashtags, and usernames. Data cleaning can be quickly done with Python. The tweet-processor library is used. With this library, we can directly use the clean method to clean the data in this library.

Case folding should also be done to change all letters to lowercase letters to improve the data for subsequent processes. This process will eliminate the difference between lowercase and uppercase letters during classification, resulting in better results. Stop word removal aims to delete words that are not important, such as unnecessary adverbs like "untuk", "pada", "ke", and "sehingga". This process is necessary because such words often do not affect the sentence and only slow down the classification process. PySastrawi Library can be used for stop word removal, as it already has a list of stop words in the Indonesian language. Other words can also be added to this list. Stemming is the process of changing affixed words into essential words. The Pysastrawi Library is used for stemming Indonesian words in this research. The last pre-processing step is tokenization, which involves converting sentences into separate individual words. This step facilitates the classification process.

3.3 Manual labeling

Manual labelling is used to determine the label of each tweet of Twitter users applying to pre-process. The complete flow of manual labelling activities is depicted in Figure 3.

First of all, the program will determine the number of max variables. If the max variable is one, set the label based on the variable. While the number of the max variable is more than one, the program will calculate the variance of the word in each max variable. The maximum number of variances will be taken as a label.

3.4 Extract features

First, for features to be used in the classification using machine learning, they will be extracted into numbers. For extract method used in this research is Bag-of-Words (BoW) and Term Frequency (TF) Inverse Document Frequency (IDF). The BoW model is the most straightforward representation of words in numbers. This model will use machine learning to make it easier to process numbers than words.

3.5 Classification

Machine learning will be used to evaluate the new algorithm that improved. The method used in this research is the decision tree. Compared to the Support Vector Model, Naive Bayes, and K-Nearest Neighbors, the decision tree has the highest accuracy [15]. Also, the decision tree is strong, has a simple structure, and is easy to implement and understand [16].

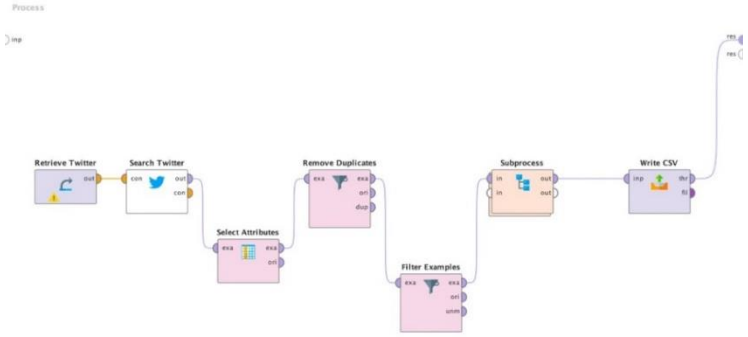


Figure 1: Data collecting and pre-processing scheme with operating the tool RapidMiner.

4 Result and discussion

This section will provide the result of classification using the Decision Tree method and will be compared with previous research using the Naïve Bayes method.

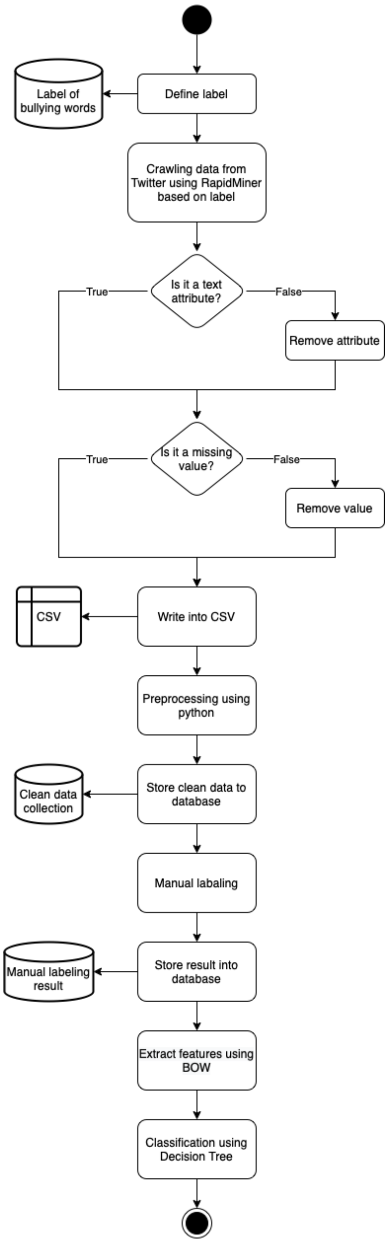


Figure 2: Data process flow

4.1 Pre-processing

Table 2 displays the raw data collected from RapidMiner, which consisted of 10,786 entries that required cleaning. The Twitter posts collected were limited to those with specific labels, and only the "text" attribute was used in this study. Other attributes, such as posting time and username, were removed during the pre-processing stage. Following pre-processing, the data became cleaner, as presented in Table 3.

Table 2: Raw data

	Text
0	Teror ”jadul”. Pembunuhan thd tiga anjing utk ...
1	Jungkook BTS punya anjing baru! #FollowLINET...
2	Penggal Kepala Anjing untuk Teror Habib Bahar,...
3	@kxiori Anjing
4	@idxn_23 Anjing Idang
…	…
10786	@mossablosslaa @AREAJULID Pentingnya akses pen...

Table 3: Raw data

	Text
0	[teror, jadul, bunuh, thd, anjing, utk, takut, ...]
1	[jungkook, bts, anjing]

2	[penggal, kepala, anjing, teror, habib, bahar, ...]
3	[anjing]
4	[anjing, idang]
…	…
10786	[akses, didik, pelosok, gin, kampung, udik]

4.2 Manual labelling

After the data is cleaned by pre-processing process, manual labelling is performed. The manual labelling will be based on the Flow in Figure 3. The result can be seen in Table 4.

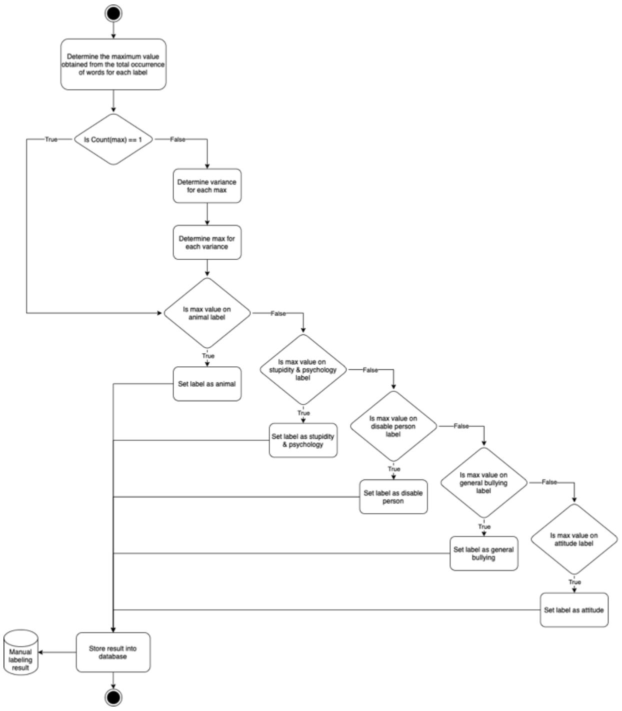
Table 4: After manual labelling

	Text	Label
0	Teror ”jadul”. Pembunuhan thd tiga anjing utk ...	Animal
1	Jungkook BTS punya anjing baru! #Fol- lowLINET...	Animal
2	Penggal Kepala Anjing untuk Teror Habib Bahar,...	Animal
3	@kxiori Anjing	Animal
4	@idxn_23 Anjing Idang	Animal
…	…	…
10786	@mossablosslaa @AREAJULID Pent- ingnya akses pen...	Psychology and Stu- pidity

4.3 Evaluation

The proposed algorithm's evaluation is measured using decision tree classification to assess the accuracy of the labels. Decision trees have been widely used in Natural Language Processing (NLP) due to their high accuracy and ease of interpretation [17]. The training and test data are divided into 80% and 20% for comparison in this method. The accuracy measure, calculated using the f-score, is 99.56%.

Figure 3: Manual labelling flow


Determine the maximum value obtained from the total occurrence of words for each label
Is Count(max) == 1
True
False
Determine variance for each max
Determine max for each variance
Is max value on animal label
False
True
Set label as animal
Is max value on stupidity & psychology label
False
True
Set label as stupidity & psychology
Is max value on disable person label
False
True
Set label as disable person
Is max value on general bullying label
False
True
Set label as general bullying
Is max value on attitude label
True
Set label as attitude
Store result into database
Manual labeling result

4.4 Comparison with previous research

In Compared to the state-of-the-art research by Febriany and Utama [5], this study introduces an additional variant feature. In this study, if there is more than one maximum value for a label, the model selects the label with the most variant words. In contrast, the previous research selected the label in the first place. This addition of variant features reduces the likelihood of unfair decisions that may have occurred in the previous study, where only the maximum value was considered.

Moreover, this study employs a decision tree as a classification method suitable for large datasets. The f-score-based accuracy of the proposed algorithm is 99.56%, which outperforms the results obtained in the study conducted by Febriany and Utama [5].

The critical difference between our study and previous research is the inclusion of a variant feature that considers the number of variant words in tweets, which can lead to a more accurate label selection. In addition, using decision trees as a classification method improves the accuracy of the proposed algorithm.

This study presents a novel approach to cyberbullying detection by combining the variant feature with decision tree classification. The proposed algorithm has the potential to be extended to other languages and platforms, providing a valuable contribution to the field of natural language processing and cyberbullying detection.

5 Conclusion

The proposed algorithm is an improved model compared to previous works, and it can handle cases where the frequency count is the maximum value of the total occurrence of the word for each label more than once. The evaluation of the model is performed using decision trees. In this study, cyberbullying is classified into five categories, namely animal, attitude, psychology and stupidity, disabled person, and general bullying, with an accuracy of 99.56%.

However, it should be noted that the model is designed only for the Indonesian language and is not applicable to other languages. In the future, it would be interesting to extend the application of this method to classify cyberbullying in English texts. Additionally, other social media platforms, such as YouTube and Instagram could also be included in the study.

References

[1] Kemp, Simon (Nov. 2021). Digital in Indonesia: All the statistics you need in 2021 - DataReportal – global digital insights. URL: https://datareportal.com/ reports/digital-2021-indonesia.

[2] Siddiqui, Shabnoor, Tajinder Singh, et al. (2016). "Social media its impact with positive and negative aspects". In: International journal of computer applications technology and research 5.2, pp. 71–75.

[3] Setiawati, Denok and Muhammad Shiddiq Al Fathoni (2020). "Relational Bullying in Re- ligious School at the Senior High School Level". In: International Joint Conference on Arts and Humanities (IJCAH 2020). Atlantis Press, pp. 170–173.

[4] Samora, Julie Balch et al. (2020). "Harassment, discrimination, and bullying in or- thopaedics: a work environment and culture survey". In: JAAOS-Journal of the American Academy of Orthopaedic Surgeons 28.24, e1097–e1104.

[5] Febriany, Ade and Ditdit Nugeraha Utama (2021). "Analysis model for identifying negative posts based on social media". In: International Journal of Emerging Technology and Advanced Engineering 11.10, pp. 104–108. DOI: 10.46338/ijetae1021_12.

[6] Mat, Tiliza Awang, Adidah Lajis, and Haidawati Nasir (2018). "Text Data Preparation in RapidMiner for Short Free Text Answer in Assisted Assessment". In: 2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA). IEEE, pp. 1–4.

[7] Vyas, Vishal and V. Uma (2018). "An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner". In: Procedia Computer Science 125. The 6th International Conference on Smart Computing and Communications, pp. 329– 335. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.12.044. URL: https://www.sciencedirect.com/science/article/pii/S1877050917328089.

[8] Chen, Caixia, Liwei Geng, and Sheng Zhou (2021). "Design and implementation of bank CRM system based on decision tree algorithm". In: Neural Computing and Applications 33.14, pp. 8237–8247.

[9] Aldino, Ahmad Ari and Heni Sulistiani (2020). "Decision Tree C4. 5 Algorithm For Tuition Aid Grant Program Classification (Case Study: Department Of Information System, Universitas Teknokrat Indonesia)". In: Edutic-Scientific Journal of Informatics Educa- tion 7.1.

[10] Shimomoto, Erica K. et al. (2018). "Text Classification Based On Word Subspace With Term-Frequency". In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. DOI: 10.1109/IJCNN.2018.8489458.

[11] Jiang, Haiyun et al. (2020). "Understanding a bag of words by conceptual labeling with prior weights". In: World Wide Web 23.4, pp. 2429–2447.

[12] Haidar, Batoul, Maroun Chamoun, and Ahmed Serhrouchni (2017). "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content". In: 2017 1st Cyber Security in Networking Conference (CSNet). IEEE, pp. 1–8.

[13] Isa, Sani Muhamad, Livia Ashianti, et al. (2017). "Cyberbullying classification using text mining". In: 2017 1st International Conference on Informatics and Computational Sciences (ICICoS). IEEE, pp. 241–246.

[14] Nurrahmi, Hani and Dade Nurjanah (2018). "Indonesian twitter cyberbullying detection using text classification and user credibility". In: 2018 International Conference on Information and Communications Technology (ICOIACT). IEEE, pp. 543–548.

[15] Dharmarajan, K et al. (2020). "Thyroid disease classification using decision tree and SVM". In: Indian Journal of Public Health Research & Development 11.3, pp. 229–233.

[16] Ghasemi, Ebrahim, Hasan Gholizadeh, and Amoussou Coffi Adoko (2020). "Evaluation of rockburst occurrence and intensity in underground structures using decision tree approach". In: Engineering with Computers 36.1, pp. 213–225.

[17] Marie-Sainte, Souad Larabi et al. (2018). "Arabic natural language processing and machine learning-based systems". In: IEEE Access 7, pp. 7011–7020.