Antisocial Behavior Monitoring Services of Indonesian Public Twitter Using Machine Learning

F A Astuti

Badan Pusat Statistik, Jakarta

*Corresponding author's e-mail: fitri.andriastuti@gmail.com

Abstract. Antisocial behavior is a personality disorder that has characteristics such as repetitive actions that violate social norms, deceit and lying, impulsiveness, irritability and aggression, reckless disregard for the safety of oneself and others, consistently irresponsible, and lack of remorse. The cause can be from various factors, including genetics, psychological conditions, interactions in the environment, and wrong parenting. The impact of antisocial behavior on social life can cause people to tend to be aggressive and take it into action by not having feelings of guilt for their actions. Thus, a monitoring of antisocial behavior disorders is needed so that it can be a warning for the public to be more concerned about the difficulties experienced by each other. The potential gained from the availability of tweet data access from the Twitter API opens up opportunities for monitoring antisocial behavior. By utilizing traditional machine learning and deep learning methods, it can be an opportunity to automate labeling on Twitter data that contains elements of antisocial behavior. Based on the description of the problems and opportunities found, this study proposes a multi-class classification monitoring service to identify public antisocial behavior on Twitter Indonesia using machine learning.

1. Introduction

Antisocial behavior is one of the indications of a personality disorder or better known as antisocial personality disorder. Antisocial behavior is included in Cluster B along with borderline, histrionic, and narcissistic personality disorder (Association, 2013). Antisocial behavior is a mental health disorder that has several characteristics that are often shown in someone who suffers from the disorder. Characteristics of antisocial behavior include repeated actions that violate social norms, deceit and lying, impulsiveness, irritability and aggression, reckless disregard for the safety of oneself and others, consistently irresponsible, and lack of remorse (Singh, 2020). In addition, there are several other characteristics such as stealing, lying, lack of remorse towards other people and living beings, irresponsible behavior, impulsive behavior, abuse of alcohol or drugs, violating the law, violating the rights of others and aggressive behavior (Nuryati and Kresnowati, 2018)). These characteristics are usually triggered by various factors causing antisocial behavior disorder.

The cause of antisocial behavior disorder is usually caused by genetic factors and psychological conditions (American Psychiatric Association, 2013). However, sometimes it can be caused by interactions in the environment, wrong parenting, low socioeconomic status, and gender causes. In one study, adults aged 18-64 years, there were about 3.3 percent of people with antisocial personality disorder of which 4.9 percent were men and 1.8 percent were women (Mental Health Foundation, 2016). From these data, it can be seen that antisocial behavior is more common in men than women.

Therefore, monitoring of antisocial behavior needs to be watched out for by considering its impact on social life.

The impact of antisocial personality disorder on social life can cause people to tend to be aggressive and take it into action by not having feelings of guilt for their actions. Perpetrators are usually referred to as sociopaths, namely people who suffer from antisocial personality disorder characterized by a lack of empathy for others, abnormal moral behavior, inability to conform to societal norms (Nuryati and Kresnowati, 2018). Perpetrators usually tend to blame others, so this personality disorder requires more attention to be treated immediately because it can harm many people. Therefore, it is necessary to monitor antisocial behavior disorders so that it can be a warning for the public to be more concerned about the difficulties experienced by each other.

The development of information technology and data storage media makes social media data a very rich source of data or commonly known as 5V, namely volume, velocity, variety, varicity, and value. For example, social media Twitter in the third quarter of 2020 had a total of 353 million users, of which 187 million active users every day also post statuses in the form of tweets, pictures, and videos (Newberry, 2021). From this data, it can be seen that social media is part of everyday life so that it has the potential to change the way of collecting information to understand society (Astuti and Nisa', 2020). The potential obtained from Twitter data and the availability of data access from the Twitter API opens up opportunities for monitoring antisocial behavior. The problem is how monitoring can be done automatically so that the public can immediately find out the statistics of tweets containing elements of antisocial behavior and can immediately take the necessary actions.

In an effort to automate behavior monitoring from Twitter social media, various machine learning methods and even deep learning can be used. The use of both traditional machine learning and deep learning can solve a variety of text, image, and video classification problems. So, this is an opportunity to automate labeling on Twitter data that contains elements of antisocial behavior.

Based on the description of the problems and opportunities found, this study proposes a multi-class classification monitoring service to identify public antisocial behavior on Twitter Indonesia using machine learning. This study focuses on processing tweet data in Indonesian and is limited to the territory of the Negara Kesatuan Republik Indonesia (NKRI). With this service, it is hoped that it will increase the awareness of the Indonesian people through monitoring the growth of antisocial behavior based on data from Twitter. This needs to be done to prevent antisocial behavior in order to reduce the negative impact it causes.

2. Background

2.1. Text Classification

Text classification automatically consists of two stages, namely feature engineering and label prediction (Singh et al., 2020). Feature engineering is the process of extracting features from the input data and its vector number representation. Some of the feature engineering techniques that are usually used for text classification are Term Frequency Inverse Document Frequency (TF-IDF), Bag-of-Words, topic modeling features, Psycholinguistic features, Sentiment lexicon features, Word n-grams, and Word Frequency (Singh et al., 2020). The next stage is label prediction where at this stage the machine learning model is trained on a benchmark data set that is extracted and annotated features are performed, which is also known as the ground truth dataset.

The challenge of automatic text classification techniques is that the text comes from humans with the ability to understand words naturally while the capabilities of computer machines are limited. Feature engineering techniques such as TF-IDF and Bag-of-Words are sometimes not very effective when dealing with problems in Natural Language Processing (NLP). This is due to the lack of semantic representation of the text corpus and the inherent dispersion problem. To overcome these shortcomings, deep learning techniques are needed that make it possible to capture not only the meaning of different words but also their interdependencies, leading to a computer understanding and context of a text.

The feature engineering techniques used in deep learning are Embedding, Word2Vec and GloVe. Both of these techniques can solve important problems in text classification such as misspellings,

synonyms, and abbreviations that are common in data collected from social media. This will make a significant performance improvement in solving text classification problems with machine learning.

2.2. Application of Machine Learning

The term machine learning was first popularized by Arthur Samuel, a computer scientist who pioneered artificial intelligence in 1959. According to Arthur Samuel, machine learning is a branch of science that gives computers the ability to learn without being explicitly programmed. Machine learning is a branch of artificial intelligence. Artificial intelligence or artificial intelligence has a very broad meaning but in general it can be understood as a computer with human-like intelligence. While ML has a more specific meaning, namely using statistical methods to make computers able to study patterns in data without needing to be programmed explicitly. Furthermore, deep learning is a branch of machine learning with artificial neural network algorithms that can learn and adapt to large amounts of data. The artificial neural network algorithm in deep learning is inspired by the structure of the human brain.

Deep learning is a computational model consisting of several processing layers to study data representations with various levels of abstraction (LeCun et al., 2015). Currently, the development of deep learning methods is growing rapidly due to the ability of computers to process deep learning algorithms. The use of deep learning algorithms can solve various problems such as classification of text, images, and videos.

2.3. Related Research Review of Antisocial Behavior

In a study conducted by Singh et al. (2020) examined the identification of antisocial attitudes from Twitter using the deep learning method. In this study, we compare four deep learning algorithms, namely CNNs, GRUs, LSTMs, and RNNs. The dataset used is 25,000 tweets based on keywords related to antisocial behavior on the Twitter platform. Most of the tweets contain sarcasm and are in the form of jokes. However, after filtering, there are only 5,504 tweets that can be processed into training data as a corpus benchmark. From the data, the tweets were classified into five classes with details of the class failure to conform to social norms as many as 1,192 tweets, irritability and aggressiveness as many as 1,238 tweets, reckless disregards for the safety as many as 804 tweets, lack of remorse as many as 868 tweets, and non-antisocial or general tweets as much as 1,402 tweets. Of these 5,504 tweets, antisocial attitudes were identified. Then these tweets are classified by the four deep learning algorithms used, namely CNNs, GRUs, LSTMs, and RNNs. The measurement results of the four algorithms, GRU shows the highest accuracy value both on the GloVe feature-set of 99.2 and on the Word2Vec feature-set of 98.60. In addition to using deep learning algorithms, this research also uses traditional machine learning algorithms such as Random Forest (RF), Decision Tree (DT), Logistics Regression (LR), and Support Vector Machine (SVM). The measurement results of these four traditional machine learning algorithms show that SVM has the highest accuracy of 94.99 on the GloVe feature-set and 96.62 on the Word2Vec feature-set. From the results of this study, it can be concluded that deep learning algorithms have better performance than traditional machine learning algorithms. This corresponds to the computational costs required by deep learning are greater than traditional machine learning.

3. Methodology

This section describes the method or design used to develop a monitoring service for the multi-class classification of Twitter Indonesia's public antisocial behavior using machine learning. The steps taken at each stage in the proposed method can be seen in Figure 1.

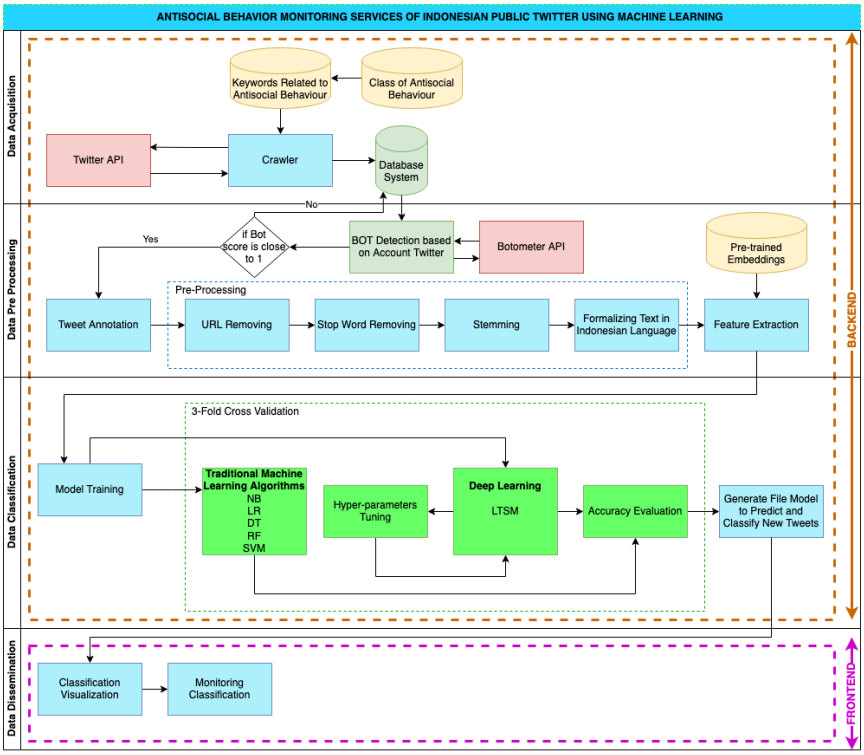

Figure 1. Antisocial Behavior Monitoring Service Architecture Design.

3.1. Data Acquisition

At this stage of data acquisition, identification of classes related to antisocial behavior was carried out. Classes are defined in five categories. First class (ClassId 1), ***Failure to conform to social norms of lawful behavior***. The examples of this class are sentences that contain elements of unlawful and illegal behavior include “I always right”, “It was not my fault”, “It's up to me to be free”, and “I will never lose”. Second class (ClassId 2), ***Irritability and aggressive***. This class contains harsh and aggressive words that are usually spoken when emotions are angry. Usually the words in this class contain elements of unlawful and illegal behavior such as "asshole, dammit, dog, bitch, dammit, asshole, and asshole". The third class (ClassId 3), ***Reckless Disregard for Safety***. This class includes reckless disregard for safety and lack of remorse like “wow you can trick this traffic violations”. Fourth class (ClassId 4), ***Lack of Remorse***. This class contains sentences with words that mean things like “I'm glad to see you cry”, and “I'm happy knowing you're sad”. Fifth class (ClassId 5), ***Non-Antisocial or General Class***. This class contains positive words such as about hope, fun, love, and affection. This class also contains tweets containing news, daily discussions, and even about business. So, this class does not contain negative words.

3.2. Data Pre Processing

At this stage, bot detection, tweet annotation, pre-processing, and feature extraction are carried out. The data preprocessing carried out includes filtering commercial tweets, url removing, stop word removing, stemming, and formalizing text in Indonesian language. At the data pre-processing stage, several tweets were handled, such as the data annotation stage and the url removing stage, stopword removing, stemming, and formalizing. Bot account detection was carried out with the API botometer [7]. At the data annotation stage, tweets containing antisocial behavior were classified using random forest. At the url removing stage, the author creates code to delete characters, usernames, retweets, urls, double spaces, and enter. The url removing stage is remove character contains http or https. The stopword removing stage is remove meaningless words. Stemming stage is parsing the form of a word into its basic word form. The formalizing stage is the stage for normalizing words that do not exist in

Indonesian, such as shortening words, using slang, spelling errors, and using inappropriate language. At this stage the process uses the REST API Pujangga [8].

3.3. Data Classification

At this stage, the modeling process of traditional machine learning algorithms such as Decision Tree (DT), Logistics Regression (LR), Random Forest (RF), Support Vector Machine (SVM) is carried out. Then the process of making a classification model is also carried out using deep learning algorithms such as Long Short-Term Memory networks (LSTMs). The model that has been made in each algorithm is then used to predict and classify the new tweets that appear. So that the monitoring results will be obtained in a time series.

3.4. Data Dissemination

This stage is the stage to visualize the classification results from tweets data that has been processed with the previous classification model. This visualization is in the form of graphs and tables. The next step is monitoring classification, which is the stage to display time series data for each class in detail. So, it is hoped that the growth of antisocial behavior in Indonesia can be seen whether it is decreasing or increasing.

4. Experiment Design and Analysis

4.1. Implementation

During crawling data from Twitter, the author uses the Twitter API V2 service. In practice, each basic user key is given a limit of 500,000 tweets and academic research users can access 10 million tweets for a full month. When developing the antisocialina project, the author uses a basic user and has used a quota of 300,403 tweets to get 165,355 tweets. In addition, there are the same request restrictions as Twitter API V1.1, which is for the past 7 days, and every 900 requests is limited to a maximum of 15 minutes.

The results of Twitter crawling on keywords that are suspected to contain elements of antisocial behavior are as shown in Table 1, in the period from April 10, 2021 to April 16, 2021, 165,355 tweets were obtained. Then deletion of Tweets is carried out by eliminating duplicate tweet fields such as retweets and tweets that have similar words by using the Levenshtein Distance score with the initial tweet that already exists in the database, if this value is closer to 0, the more similar the words are. So, from these results, 85,654 tweets were deleted, and 79,701 tweets were obtained. Then filtering tweets that contain indications of antisocial behavior or not by using the random forest algorithm from the training data that has been made previously, so that a total of 33,768 tweets are processed. After that, a sample was carried out to create training data from the classification of the five antisocial classes as many as 1,251 tweets.

Table 1. List Keyword and Class for Crawling.

Keyword	ClassId	Keyword	ClassId	Keyword	ClassId	Keyword	ClassId
kebebasanku	1	Anying	2	Bunuh saja tanpa	3	Senang bikin sedih	4
bukan salahku	1	Njir	2	Bahaya tapi bisa	3	Senang gagal	4
selalu benar	1	Bajingan	2	Selamat menikmati karmamu	3	Suka gagal	4
langgar hukum	1	Jalang	2	Asik bahaya	4	Suka menderita	4
bodo amat	1	Lonte	2	Suka bahaya	4	Suka menyesal	4
bukan urusanku	1	njeng	2	Senang bikin susah	4	Semoga doa terkabul	5

Keyword	ClassId	Keyword	ClassId	Keyword	ClassId	Keyword	ClassId
Brengsek	2	Asik bisa tanpa	3	Bahagia bikin susah	4	Syukur	5
Sialan	2	Suka bisa tanpa	3	Suka bikin susah	4		
Anjing	2	Cepat jadi tanpa	3	Suka bikin sedih	4		

At the data pre-processing stage, several tweets were handled, such as the data annotation stage, the url removing stage, stopword removing, stemming, and formalizing. The results of the pre-processing process become training data that is used as the basis for making models. The data that has been created for training data in this study have been uploaded by the author on the Kaggle platform at the address https://www.kaggle.com/fitriandri/antisocial-behaviour-public-twitter-indonesia.
After that, the tweet data labeling stage was carried out into five classes with a sample of 1,251 tweets. The Table 1 shows the composition of the data labeling and the composition of the sample for each class. In making labeling datasets or training data, entering a sentence into a predetermined class is based on the intuition of the dataset maker. In the context of the classification of antisocial behavior by using sentences, it is part of social science which is also very subjective in nature.

Table 1. Tweet Sample Composition for Each Class.

Class	Count
Failure to conform social norms of lawful behavior	272
Irritability and aggressive	132
Reckless Disregard for Safety	39
Lack of Remorse	22
Non-Antisocial or General Class	786

4.2 Classification Model Evaluation Results

In the classification stage, tweets are classified into five predetermined classes. The models used at this stage include machine learning algorithms such as Naïve Bayes (NB), Logistics Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). In addition, classification is carried out using a deep learning model using Long Short-Term Memory networks (LSTM). The features text extraction techniques used are TF-IDF and Embedding Word2Vec.

Word embeddings is a process in NLP in converting words in the form of alphanumeric characters into vector form, one of which is using the Word2Vec model. To build the Word2Vec model, the Word Embedding process can be done with the help of the Gensim library. In this study using vector size of 100 dimensions from pretrained wikipedia data.

The architectural design of the LSTM consists of several steps. the first step is to do a random split train and test data (90:10). Next, create an LSTM model with 6 Layers (Input Output Layer, Dropout1, Hidden Layer, Dropout2, Dense Layer, and Activation Layer). Then, setting the hyperparameter tuning using Adam Optimizer. After that, the fitting model configuration is carried out, which uses 2048 Batch Size, 10 Epoch, and 0.1 Validation Split.

Based on the training data and the classification model that has been made previously, the measurement evaluation results are obtained as shown in Table 2. This study uses 3-fold cross-validation for accuracy evaluation where the overall dataset is divided into training data and testing data. In this research, kfold class from sklearn library is used with parameter 3 n_splits. Based on the results in Table 2 the random forest algorithm has the best accuracy so that it is used as the main model to predict other tweets.

Table 2. Machine Learning Model Measurement Evaluation Results.

Model	Feature-Set	Accuracy	Precision	Recall	F1_score	Support
Naïve Bayes	TF-IDF	0,65	0,64	0,67	0,56	1001
Logistics Regression	TF-IDF	0,73	0,73	0,73	0,72	376
Decision Tree	TF-IDF	0,75	0,75	0,75	0,74	251
Random Forest	TF-IDF	0,76	0,71	0,76	0,73	313
Support Vector Machine	TF-IDF	0,72	0,71	0,72	0,67	1001
Long Short-Term Memory	Word2Vec	0,67	0,44	0,67	0,53	126

4.3. Prediction Results of Antisocial Behavior Classification

Based on the data dissemination stage, the following are the results of the model's prediction of 7,409 other tweets selected randomly from 33,768 tweet as mentioned in data acquisition before, as shown in Table 3.

Table 3. Prediction Results from Tweet Testing.

Class	Count
Failure to conform social norms of lawful behavior	242
Irritability and aggressive	2,373
Non-Antisocial or General Class	4,794

4.4. Implementation of the Antisocialina Monitoring Service API

At this stage, the implementation of the creation of several endpoints for monitoring antisocial behavior from the Indonesian Twitter public is carried out. Deploy the implementation of this monitoring service at https://antisocialinadev.herokuapp.com/docs. Figure 2 is a list of endpoints created to monitor antisocial behavior from Twitter media.

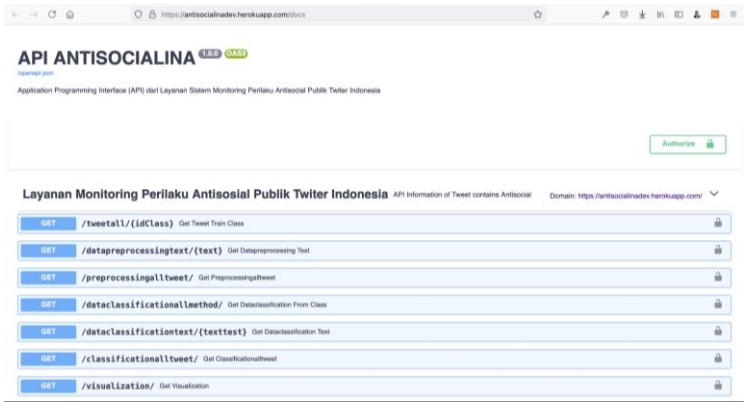

Figure 2. List of Antisocial Behavior Monitoring Service Implementation Endpoints.

4.5. Implementation of Antisocial Behavior Monitoring on the Antisocialina Application

This stage is the dissemination stage from the data acquisition stage to data classification in the form of tables and time series line graphs. As shown in Figure 3 is one of the frontends of this antisocial behavior monitoring service. More complete prediction results can be seen on the frontend with the address https://antisocialidmonitor.herokuapp.com/.

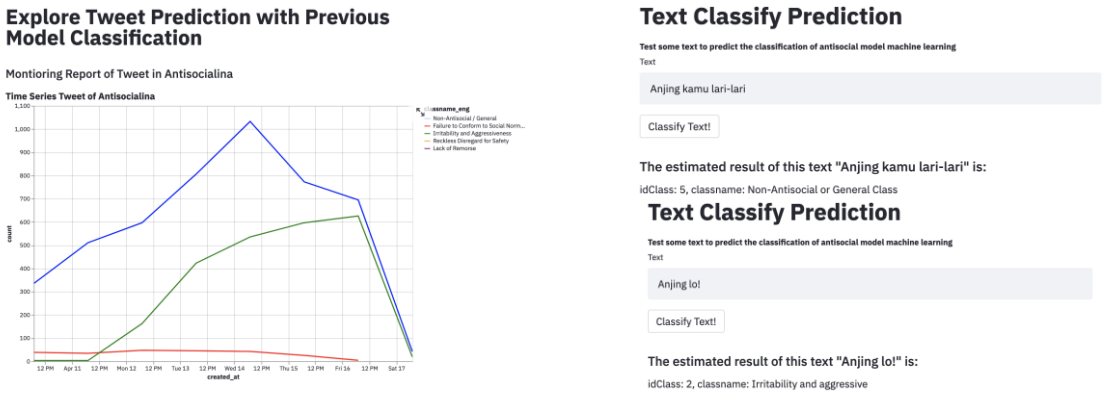

Figure 3. Predicted Result Tweet Component.

5. Conclusion

Based on the results and discussion obtained in this study, it can be concluded that the labeling process carried out in this study resulted in the random forest algorithm as the algorithm with the best and same accuracy with an accuracy of 0.76. It turns out that in the data labeling carried out by the author, traditional machine learning algorithms have better performance than deep learning such as LSTM. This is thought to be caused by a sample that may be biased because the authors still have not found the best composition for labeling each class as shown in Table 1 the number of tweets in classes 1,2, and 5 dominates over classes 3 and 4. This is in accordance with the reality in when fetching tweets with less keyword crawlers for grades 3 and 4. So this causes the labeling composition to be less balanced, each class should have the same number. However, the model built using the random forest model with an accuracy of 0.76 was able to predict tweets into the antisocial behavior class.

Overall, the model can be used to predict tweet testing with a test sample of 7,409. However, with not very high accuracy, it causes a lot of misclassifications of predictions for some classes. Therefore, the author suggests creating a more balanced and increased dataset class. In addition, the author suggests adding a feature that not only uses TF-IDF or Embedding from the hard library but by using Word2Vec or GloVe for better accuracy.

References

[1] American Psychiatric Association, (2013): Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Washington, DC, USA: American Psychiatric Pub.

[2] Astuti, F., A. dan Nisa', N.A. (2020): Harnessing Social Media Data to Measuring Mental Health Statistics, 2020 Asia–Pacific Statistics Week, https://www.unescap.org/sites/default/files/APS2020/49_Harnessing_Social_Media_Data_to_Measuring_Mental_Health_Statistics_Indonesia.pdf.

[3] LeCun, Y., Bengio, Y., Hinton, G. (2015): Deep learning, Nature, 521, 436-444, doi: 10.1038/nature14539.

[4] Mental Health Foundation (2016): Fundamental Facts about Mental Health 2016, https://www.mentalhealth.org.uk/sites/default/files/fundamental-facts-about-mental-health-2016.pdf.

[5] Nuryati dan Kresnowati, L. (2018): Klasifikasi dan Kodefikasi Penyakit dan Masalah Terkait III Anatomi, Fisiologi, Patologi, Terminologi Medis dan Tindakan pada Sistem Panca Indra, Saraf, dan Mental, Kementrian Kesehatan Republik Indonesia, pp. 36, http://bppsdmk.kemkes.go.id/pusdiksdmk/wp-content/uploads/2018/09/Klasifikasi-Kodefikasi-Penyakit-Masalah-Terkait-III_SC.pdf.

[6] Singh, R. et al. (2020): Deep Learning for Multi-Class Antisocial Behavior Identification From Twitter, IEEE Access, vol. 8, pp. 194027-194044, 2020, doi: 10.1109/ACCESS.2020.3030621.

[7] Davis, C. A. et al. (2016): BotOrNot: A system to evaluate social bots, Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273-274.

[8] Purwarianti, A. et al. (2016): InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification, 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), pp. 1-5, doi: 10.1109/ICAICTA.2016.7803103.