Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2025 (43) 4** : 1–14.
ISSN-0037-8712
doi:10.5269/bspm.78695

Detection of Bahasa Cyberbullying Speech Using Large-scale N-Gram Machine Learning Models with Increased Document-Terms Probability

Yudi Setiawan*, Endina Putri Purwandari, Andang Wijanarko, Yusran Putra Panca, Ferzha Putra Utama

ABSTRACT: A rising number of bullying incidents, whether between people or groups (cyberbullying), can be attributed to the proliferation of social media technologies and sharing websites. One difficulty in identifying cyberbullying in Bahasa is that words can have more than one meaning when combined with another, making them ambiguous or even negative. In this article, we look at how to increase the probability value of document-terms in a machine learning model to achieve high classification accuracy in the detection of Bahasa cyberbullying, which features a wide range of meanings, word spellings, and meaning shifts on social networking platforms. In addition, a language model with sequential sequences of n-words to capture patterns and statistics in the text data (Large-scale N-Gram) is applied throughout the detection phase to categorize texts based on the cyberbullying corpus created during training and testing. Our research shows that the accuracy of Indonesian cyberbullying detection may be greatly enhanced by collecting trends and boosting the probability value of document-terms.

Key Words: Cyberbullying, document classification, large-scale n-gram, machine learning model.

Contents

1. Introduction

The proliferation of technology that enable social media has resulted in a rise in the amount of attention that is dedicated to the study of cyberbullying. (A New Bottle to Hold an Old Wine) (Chan et al., 2021; Q. Li, 2007), cyberbullying is a well-known example of a traditional action dilemma in which one person poses a danger to another. One might say that it is an extension of the traditional form of bullying, and that it is made feasible by the capability of social media and electronic communication technology to cover enormous distances. Cyberbullying is often not a physical form of harassment, but it can involve the leaving of digital traces, such as images and videos, which can have a major influence

* Corresponding author.
2010 *Mathematics Subject Classification:* 68T50.
Submitted August 30, 2025. Published November 01, 2025

on the victim's social life (Sheldon et al., 2019). This is according to Sheldon et al. (2019), who state that cyberbullying is frequently not a physical form of harassment.

The term "cyberbullying" refers to any negative behaviour that takes place in the digital world, such as forms of harassment, threatening, slander (false information), brief insults (flaming), and the activity of recording someone who is being bullied. These behaviours can be carried out briefly or repeatedly by individuals or groups aimed at victims (Balakrishnan et al., 2020; T. Mahlangu & C. Tu, 2019). The act of cyberbullying is a change in form from traditional bullying that occurs face to face to interactions that occur in the digital world, generally through social media applications (López-Vizcaíno et al., 2021; W. M. Al-Rahmi, N. Yahaya, M. M. Alamri, N. A. Aljarboa, Y. B. Kamin, & F. A. Moafa, 2019).

The world of education is rife with instances of cyberbullying, which are often carried out by students with the intention of upsetting and threatening other students (S. Salawu et al., 2020; W. M. Al-Rahmi, N. Yahaya, M. M. Alamri, N. A. Aljarboa, Y. B. Kamin, & M. S. B. Saud, 2019). Research has shown that bullying in schools has been a significant and pervasive problem for several decades in a variety of nations (Q. Li, 2007; Noviantho et al., 2017. Students with unequal power engage in cyberbullying towards other students, and this behaviour is carried out repeatedly. This has a significant negative impact on the mental and physical health of students, and it has even led to death threats in a number of nations (Barlett, 2019). Because it may be difficult for socially anxious adolescents to begin social contacts, be accepted in bigger peer groups, and build close friendships (I. Ting et al., 2017), social anxiety is a major concern for children and adolescents. This is a problem because socially anxious adolescents may find it difficult to initiate social relationships. As a result of the lack of role that teachers or school administrators play in knowing and recognising cyberbullying acts committed by students to other students, it also develops as a result of the fact that schools sometimes conduct many investigations tending to cyberbullying with the goal of surveying the similarities (V. Banerjee et al., 2019).

The challenge that arises with digesting text is identifying what the text actually means. This is because there are inconsistent linguistic variances in text writing, differences in writing styles that are not common and standardised, and the usage of words that carry ambiguous meanings, which results in high processing needs in order to get the meaning of the contextual text accurate. According to Baggini and Fosl (Baggini & Fosl, 2010), a deductive argument is considered or presented as valid if the conclusion follows the premises. If the conclusion does not follow the premises, then the argument is considered or shown as invalid. There are still prospects for development in feature extraction research using large-scale N-Gram to carry out the text classification procedure, particularly in text classification for Bahasa cyberbullying detection. This is due to the fact that Bahasa cyberbullying detection patterns have various ways of detecting the meaning of sentences and phrases, in addition to having rules for sentence structure that are distinct from those found in other Latin texts. Therefore, the purpose of this article is to create a dataset, a corpus, and a training procedure for cyberbullying and non-cyberbullying in Bahasa Indonesia in order to be able to handle syntactic difficulties in the case study of Bahasa cyberbullying categorization.

2. Contribution

This research investigates the identification of Bahasa cyberbullying, which may have a number of different connotations depending on context, word spelling, the mixing of languages in communication, and changes in the meaning of words used in the language of communication in social media. Due to the fact that this causes the potential of varied perceptions to interpret the delivery of information or speech, as well as the use of biased phrases, it becomes difficult to carry out the process of detecting cyberbullying. Therefore, it is vital to examine this research by putting into action a model of machine learning in order to carry out the learning process and classify instances of cyberbullying in Indonesia. In both the cyberbullying training process and the non-cyberbullying training process, a feature extraction procedure is carried out. This is done so that the accuracy of the cyberbullying categorization may be improved. Implementing Large-Scale N-Gram is what is done throughout the process of feature extraction.

This allows for the determination of the frequency of word occurrence based on the sequence of the words that make up the sentence or document. A probabilistic method is applied to the calculation of the frequency with which each word appears before arriving at a conclusion on the word's likelihood. The

contributions of this research are:
1. Producing a study of Indonesian cyberbullying patterns based on the syntax of utterances/comments based on the frequency of occurrence of terms. This is one of the contributions to the domain of algorithm adaptation of this research.
2. Constructing a model for the identification of cyberbullying in Bahasa by using machine learning as a kind of learning and classification.
3. Create an updated cyberbullying corpus in Indonesian that takes into account current word usage patterns.
4. Improve the effectiveness of the cyberbullying detection procedure by including Large-Scale N-gram into the process of feature extraction.Large-Scale N-gram into the process of feature extraction.

3. Motivation

A hostile act committed on social media with the intent to harass, humiliate, or ridicule another user is an example of cyberbullying. The motivation may be broken down into two categories: internal motivation, which originates from the individual themselves, and external motivation, which originates from the environment around them. The internal motivation may be broken down into five categories, which are as follows: (1) Bullying occurs in the real world, which then leads to it occurring on social media since it is the simplest method to bully someone online. (2) There is an imbalance of power, meaning that what the cyberbully does in terms of mental and social strength feels stronger than what the victim accomplishes. (3) The belief that they have everything necessary to do anything they want to the victim in order to hurt them. (4) Causing the sufferer emotional anguish to the point that they feel the need to change schools or maybe stop attending to school altogether. (5) If the victim is unaware of the perpetrator's identity, it will be much simpler for the perpetrator to engage in cyberbullying.

In addition, the use of language or dialect as a means of communication on social media results in a variety in spelling and writing, which can result in arguments and threats. The use of terms that are not standardized or that have a meaning that is biased is another obstacle that makes it difficult to identify instances of cyberbullying in today's world. Based on these factors, it is clear that research on the identification of cyberbullying currently faces obstacles that either continue to expand or remain unchanged. It is necessary to implement an information technology strategy in order to be able to identify instances of cyberbullying on social media platforms, and it is anticipated that this strategy will be able to stop discussions, comments, and content that can lead to cyberbullying from being published on social media platforms.

4. Research Workflow

classification method based on machine learning is what we propose for the identification of cyberbullying in Bahasa documents. The purpose of using machine learning models to the process of detecting cyberbullying is to identify patterns in the data and construct statistical models that can be used to make choices or predictions on fresh data that has never been seen before. This new data has not been seen before. Therefore, the purpose of this research is to develop methods and algorithms that will enable computers to learn from data and experience without being explicitly programmed precisely, and to make predictions or make conclusions based on patterns or trends that can be detected in the data pertaining to the subject of cyberbullying. The stages of this investigation are shown forth in Figure 1, as is appropriate.

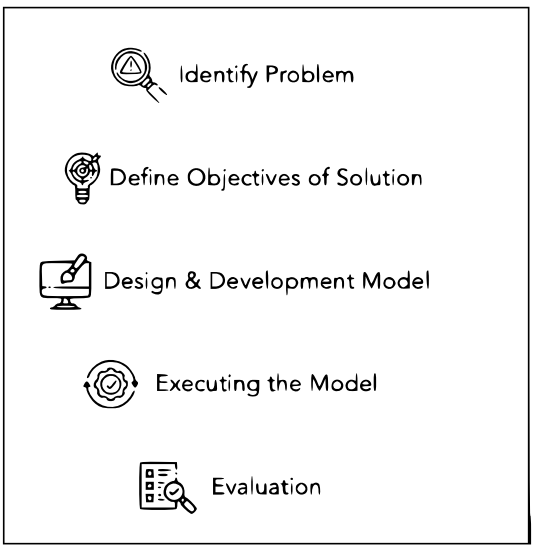


Figure 1: Research Workflow

A lack of consensus on what exactly defines cyberbullying is one of the most prevalent challenges encountered when attempting to compile statistics on the phenomenon of cyberbullying. Some people may perceive certain behaviors to be cyberbullying, while others would disagree with that assessment. A lack of consistency on definitions might make it difficult to identify and record consistent occurrences of cyberbullying in the dataset. This can lead to issues. Therefore, the first step in the research stage is to establish the research problem, the purpose of which is to define the nature as well as the definition of cyberbullying within the context of the book.

In order to create a dataset on cyberbullying, one of the challenges that must be overcome is the creation of a dataset that is representative of the whole. In this scenario, it is vital to make certain that the dataset encompasses textual forms of cyberbullying, such as bullying, harassment, disclosure of personal information, or defamation. These are the sorts of behaviors that fall under the umbrella of "cyberbullying." The constructed dataset does not include information on the gender, age, background, or any other demographic features of individuals (persecutors) or public accounts that are involved in incidents involving cyberbullying.

Building a machine learning system is how the process of detecting cyberbullying is carried out, according to the design and model for the process of detecting cyberbullying. This research creates a classification training process by utilizing a variety of algorithms and supervised learning approaches. The purpose of this process is to make comparisons and evaluate the accuracy of classification patterns that generate an accurate level of cyberbullying detection. The testing procedure is carried out by making use of the dataset that has been constructed in order to assess the accuracy and flaws that arise from the cyberbullying detection model and the dataset that has been constructed in order to carry out an evaluation of the procedure that is used for conducting research.

5. Creating a Bahasa Dataset and Corpus of Cyberbullying

esearchers and developers are able to perform analysis and train machine learning models to detect and identify patterns of cyberbullying behaviour when they have access to a sizeable and representative dataset on cyberbullying in Bahasa. This can aid in the development of efficient methods and procedures to protect persons from the hazards of cyberbullying, as well as give the essential assistance and resources to those who are victimized by it.

This information may also be utilized to establish better rules and regulations addressing protection against cyberbullying in environments where Bahasa is spoken as a primary language. The information that was gained from the study of the dataset can be helpful in gaining a better understanding of the

variables that affect cyberbullying, the behavior patterns of those who engage in it, and the effect it has on those who are bullied. Guidelines and regulations that are produced based on this knowledge can improve the process of early identification and recognition of cyberbullying activities that occur on social media. The goal of this process is to hopefully prevent incidences of cyberbullying and give support to victims of cyberbullying.

Proses pembuatan dataset cyberbullying Indonesia, dilakukan dengan beberapa tahapan untuk menghasilkan koleksi data yang kompleks dan lengkap, sehingga menggambarkan kondisi riil cyberbullying yang terjadi. Adapun tahapan-tahapan dalam pembuatan dataset dan corpus cyberbullying Bahasa Indonesia ditunjukkan pada Gambar 2. Proses pembuatan dataset cyberbullying Bahasa dilakukan dengan melakukan tahapan; pengumpulan data, proses pelabelan dan anotasi, ekstraksi data, preprocessing dan pembersihan data, pengorganisasian data, dan pembuatan korpus.

In order to develop a complicated and comprehensive data collection, the process of producing a Bahasa cyberbullying dataset was carried out in various stages. This enabled the collection to accurately describe the actual conditions under which cyberbullying occurs. Figure 2 illustrates the process that was followed in order to compile the Bahasa cyberbullying dataset and corpus. The stages of data collecting, labeling and annotation procedure, data extraction, preprocessing and data cleaning, data organization, and corpus construction are all parts of the process of producing a Bahasa cyberbullying dataset. This process is conducted out by following the steps.

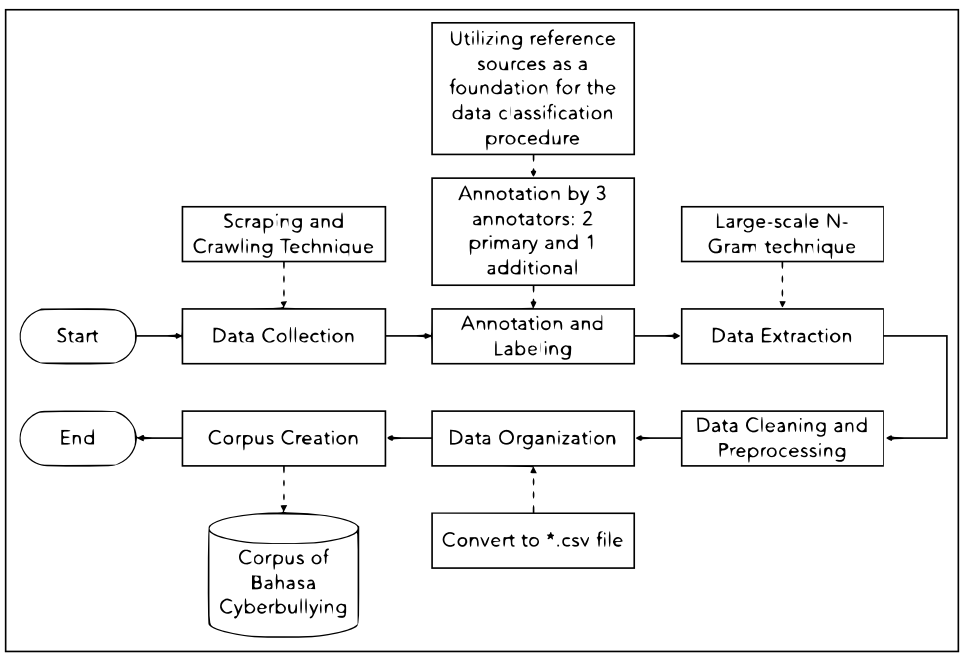


Figure 2: Workflow of Creating Bahasa Dataset and Corpus

Real-time data collecting is an additional challenge that must be overcome while developing a dataset on cyberbullying. Because incidents of cyberbullying can occur suddenly and on a variety of platforms, it can be challenging to gather and document instances that are true and reflect current trends in the field. The scraping and crawling processes are utilized to complete the data collecting procedure. Scraping techniques are used to gather datasets from the YouTube platform. YouTube gives application programming interface (API) access to Google Development, which enables third parties to use the platform for the purpose of application development. Crawling techniques are utilized by the author in order to collect datasets from the platforms of Instagram and Twitter. Bot is used in place of a search engine in order to obtain data based on links and keywords. The following categories are included in the cyberbullying search subjects: racism themes; cyberbullying in political comments or conversations; social topics; viral

or fascinating things today; and interesting things today.

In this line of study, one of the challenges is correctly labeling the data in the cyberbullying datasets. A comprehensive knowledge of the problems must be had by trained annotators in order for them to correctly and consistently categorize occurrences of cyberbullying within the dataset. Inconsistent or erroneous labeling can be the consequence of a lack of comprehension or variations in judgment, both of which have the potential to lower the quality of the dataset and its overall usefulness. Annotators are the ones that carry out the labeling process. These individuals have an expertise and knowledge of the particular domain or activities that need to be carried out in order to offer correct and appropriate labels on the data that has been provided. Annotators are responsible for a variety of data processing activities, including the assignment of class labels to the data, the labeling of particular entities or objects within the data, the addition of metadata, and other similar responsibilities.

When labeling each text or document from the data collection, you will require the assistance of two annotators. In the event that the labels that are assigned to the text or document by the two annotators are not the same, the text or document will be examined by a third annotator, who will then provide a label to the text or document.

During this stage of the study, the Large-Scale N-Gram model is the one that is utilized for the feature extraction process. The Large-Scale N-Gram model is a sort of language model that captures patterns and statistics in text data by making use of n-grams. N-grams are continuous sequences of n elements (often words or characters), and the model makes use of these sequences. Large-scale n-gram models are sometimes referred to by their alternate name, large-scale n-gram models. It is called "large-scale" because it is used to execute operations on a very big corpus of text, which is often a sizeable collection of documents or an enormous data set. The term "corpus" refers to the entirety of the text that is being processed.

The value of n in the n-gram model is what decides how many elements are included in the sequence that is being analyzed. For instance, the bigram model has a n value of two, which means that it examines successive pairs of words. On the other hand, the n value in the trigram model is equal to three, which means that it analyzes triplet words as well as subsequent word combinations. The model is able to determine the likelihood of a certain word appearing in a specific setting by first calculating the frequency of each n-gram and then calculating the probability of the word given n-1 words that came before it. In this way, the model is able to assess the likelihood of a specific word occurring in a specific context. The method of training large-scale n-gram models often entails training on massive datasets that contain billions or even trillions of words. This is done so as part of the training process. These models make use of the volume of data to identify complicated language patterns, which enables them to contribute to the production of material that is coherent and appropriate to the context in which it is presented.

During the preprocessing stage, three steps are carried out. These steps are the folding of lowercase letters, the removal of regular expressions, and the removal of punctuation words. For the first three steps of the preprocessing, the Sastrawi Library, which is a module that is owned by the Phython library, was utilized. This library is responsible for reducing inflected words in Bahasa to their standard form or in compliance with the standards of the Indonesian Dictionary.

Case folding is the process of transforming text into a common case format, either lowercase or uppercase, in order to make it easier to compare and edit text. This may be done with either uppercase or lowercase text. In this particular setting, it is essential to differentiate between the processes of case folding and case conversion. Case folding aims to achieve case insensitivity and uniformity, in contrast to case conversion, which involves altering the case of text according to particular demands (such as changing all letters to uppercase or lowercase), case folding includes changing the case of text. In spite of the fact that the regular expression removal stage works toward the elimination of certain patterns through the use of regular expressions, the regular expression removing stage makes use of the following steps:

1. Import regular expression module in programming language.
2. Specify the regular expression pattern that matches the specific pattern or patterns you want to remove from the text.
3. Apply the regular expression pattern to the text data, replacing the matching pattern with the desired

empty or replacement string.

Punctuation words like "a", "an", "the", "and", "or", "but", etc. are often omitted from extracted text during the preprocessing step of punction word removal. Using regular expression patterns, you may swap out punctuation words in the text data with null strings or new terms.

The dataset that is generated as a consequence of the preparation step will be saved in the *.csv file format. This is done so that the dataset may be processed for the purposes of corpus generation, cyberbullying / non-cyberbullying data training, and data testing procedures. Building a model in the computer language Python is done in order to carry out the processes of training, creating a corpus, and testing for the detection of cyberbullying or the absence of cyberbullying.

A huge collection of textual data is called a corpus, and it is created by merging all of the text data into a single coherent unit. The Term Frequency- Inverse Document Frequency (TF-IDF) feature extraction approach is used in this study to create an Indonesian Cyberbullying Corpus. This method is used to compile the texts that make up the corpus. Calculating the frequency of occurrence of each word in each page is one method for collecting data that has been preprocessed. Count the number of times that each word in a document appears by going through the document word by word. After then, carry on with the process of calculating IDF for each individual word in the corpus. The logarithm of the number of documents in the corpus, divided by the number of documents that include the term, is the standard formula for computing IDF. After you have determined the TF and IDF values for every word in the document and the corpus, use equation 1 to determine the TF-IDF value for every word in each document.

$$TF - IDF = TF * IDF \tag{5.1}$$

Develop a corpus using each document's TF-IDF values. The corpus may be represented as a matrix with rows representing documents and columns representing words. The matrix contains the TF-IDF values for each document word. After these procedures, TF-IDF feature extraction may be utilized for text analysis, document categorization, and information retrieval.

6. Detecting Methods and Evaluation Bahasa Cyberbullying

he cyberbullying document classification technique uses the n-Gram approach to TF-IDF feature extraction to find the best pattern for each n-Gram based on word occurrence. Several word occurrence variables are tested to find the best TF-IDF feature extraction pattern with the n-Gram method. Traditional TF-IDF algorithms are modified or combined to increase search efficiency and document word distribution (Arroyo-Fernández et al., 2019). One issue with this dictionary-based method is that it ignores sentence word frequency. Probabilistic frequency computation can calculate word probabilities.

In the training dataset, there are numerous options for calculating the sum of all terms, including;
1. Unigram approach (n=1). A word count that generates probabilities for each word in a sentence or phrase. Unigram treats every word separately. It does not consider the relationship between words in context.
2. Bigram Approach (n=2). The bigram approach involves linking two words or terms to the preceding one. This method obtains document training information every two syllables. Therefore, the dictionary generates a pair of words and their frequencies based on a single term.
3. Trigram Approach (n = 3). The Trigram approach is similar to the other approaches, but it performs statistical summation calculations on three adjacent words/syllables.
Beyond the Trigram method, n-Grams can be enlarged. This provides context because n is huge in many cases. The following feature extraction example uses n-Gram:

The original utterance: **"I reside in Bengaluru"**

Then, the n-Gram method of extraction makes a set of words that are kept in the array and shown in Table 1.

Table 1: Produce Data Results Using the n-Gram Method

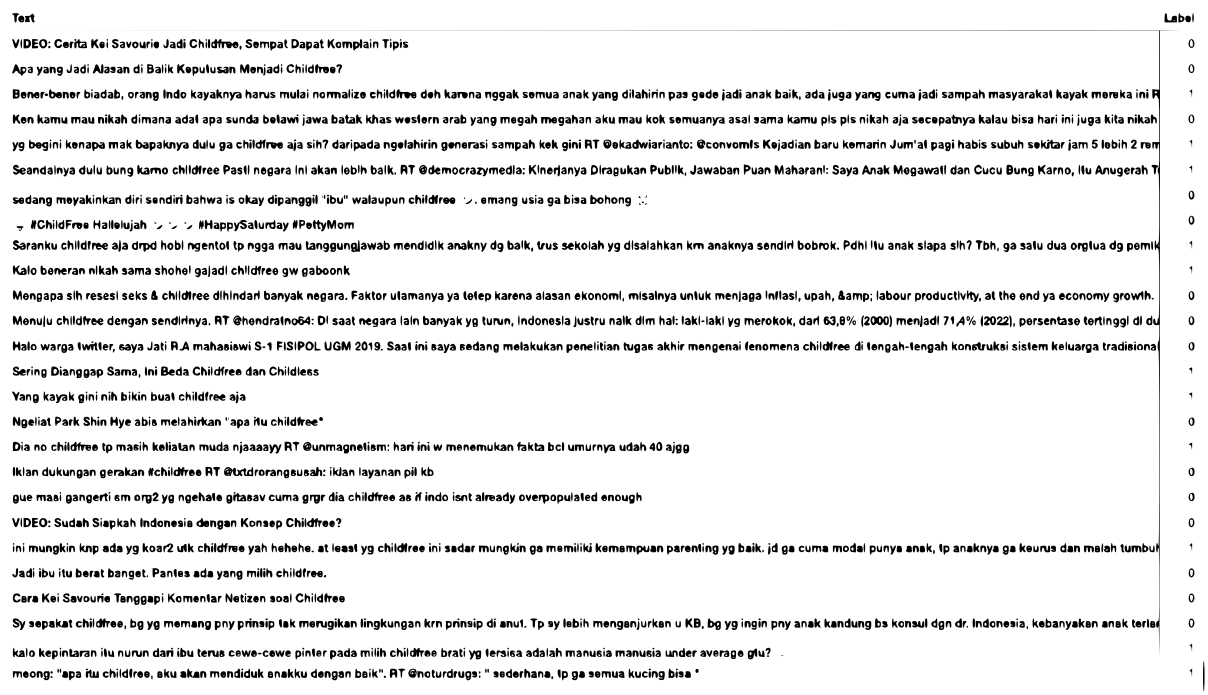
n-Gram type	n	Generate of n-Gram Method	Vector Count
Unigram	1	"I", "reside", "in", "Bengaluru"	4
Bigram	2	"I reside", "reside in", "in Bengaluru"	3
Trigram	3	"I reside in", "reside in Bengaluru"	2

7. Result-Bahasa Cyberbullying Dataset

hen applied to text, feature extraction techniques can bring about the availability of datasets that can be used for a variety of applications. Collecting text data from social media platforms, which is subsequently extracted, can assist academics in capturing patterns of user behavior on social media by studying speech and remark trends. This can be accomplished through the process of pattern recognition.

This study generates an Indonesian cyberbullying dataset by crawling and scraping data from social media sites including Twitter, Instagram, and Youtube. The dataset contains information on incidents of cyberbullying in Indonesia. The materials that were collected cover a wide range of themes and concerns, some of which are political issues, prominent people, racism, culture shock, and taboo talks. The data that was obtained has been tagged (cyberbullying label '1', non-cyberbullying label '0') by annotators who are bound by Indonesian legal limitations. "Law Number 11 of 2008 concerning Electronic Information and Transactions and its amendments".

The dataset on cyberbullying in Indonesia that was produced as a consequence contains 66,818 data recordings, 6,329 data records that are categorized as cyberbullying, and 60,489 data records that are identified as not being cyberbullying. Figure 3 presents the results of the compilation of the data set on Bahasa cyberbullying.



Text	Label
VIDEO: Cerita Kei Savourie Jadi Childfree, Sempat Dapat Komplain Tipis	0
Apa yang Jadi Alasan di Balik Keputusan Menjadi Childfree?	0
Bener-bener biadab, orang Indo kayaknya harus mulai normalize childfree deh karena nggak semua anak yang dilahirin pas gede jadi anak baik, ada juga yang cuma jadi sampah masyarakat kayak mereka ini R	1
Ken kamu mau nikah dimana adat apa sunda betawi jawa batak khas western arab yang megah megahan aku mau kok semuanya asal sama kamu pls pls nikah aja secepatnya kalau bisa hari ini juga kita nikah	0
yg begini kenapa mak bapaknya dulu ga childfree aja sih? daripada ngelahirin generasi sampah kek gini RT @ekadwiarianto: @convomls Kejadian baru kemarin Jum'at pagi habis subuh sekitar jam 5 lebih 2 ren	1
Seandainya dulu bung karno childfree Pasti negara ini akan lebih baik. RT @democrazymedia: Kinerjanya Diragukan Publik, Jawaban Puan Maharani: Saya Anak Megawati dan Cucu Bung Karno, Itu Anugerah T	1
sedang meyakinkan diri sendiri bahwa is okay dipanggil "ibu" walaupun childfree :). emang usia ga bisa bohong ::	0
#ChildFree Hallelujah #HappySaturday #PettyMom	0
Saranku childfree aja drpd hobi ngentot tp ngga mau tanggungjawab mendidik anakny dg baik, trus sekolah yg disalahkan krn anaknya sendiri bobrok. Pdhl itu anak siapa sih? Tbh, ga satu dua orgtua dg pemik	1
Kalo beneran nikah sama shohel gajadi childfree gw gaboonk	1
Mengapa sih resesi seks & childfree dihindari banyak negara. Faktor utamanya ya tetep karena alasan ekonomi, misalnya untuk menjaga inflasi, upah, & labour productivity, at the end ya economy growth.	0
Menuju childfree dengan sendirinya. RT @hendratno64: DI saat negara lain banyak yg turun, Indonesia justru naik dlm hal: laki-laki yg merokok, dari 63,8% (2000) menjadi 71,4% (2022), persentase tertinggi di du	0
Halo warga twitter, saya Jati R.A mahasiswi S-1 FISIPOL UGM 2019. Saat ini saya sedang melakukan penelitian tugas akhir mengenai fenomena childfree di tengah-tengah konstruksi sistem keluarga tradisiona	0
Sering Dianggap Sama, Ini Beda Childfree dan Childless	1
Yang kayak gini nih bikin buat childfree aja	1
Ngeliat Park Shin Hye abis melahirkan "apa itu childfree"	0
Dia no childfree tp masih keliatan muda njaaaayy RT @unmagnetism: hari ini w menemukan fakta bcl umurnya udah 40 ajgg	1
Iklan dukungan gerakan #childfree RT @txtdrorangsusah: iklan layanan pil kb	0
gue masi gangerti sm org2 yg ngehate gitasav cuma grgr dia childfree as if indo isnt already overpopulated enough	0
VIDEO: Sudah Siapkah Indonesia dengan Konsep Childfree?	0
ini mungkin knp ada yg koar2 utk childfree yah hehehe. at least yg childfree ini sadar mungkin ga memiliki kemampuan parenting yg baik. jd ga cuma modal punya anak, tp anaknya ga keurus dan malah tumbuh	1
Jadi ibu itu berat banget. Pantes ada yang milih childfree.	0
Cara Kei Savourie Tanggapi Komentar Netizen soal Childfree	0
Sy sepakat childfree, bg yg memang pny prinsip tak merugikan lingkungan krn prinsip di anut. Tp sy lebih menganjurkan u KB, bg yg ingin pny anak kandung bs konsul dgn dr. Indonesia, kebanyakan anak terla	0
kalo kepintaran itu turun dari ibu terus cewe-cewe pinter pada milih childfree brati yg tersisa adalah manusia manusia under average gtu?	1
meong: "apa itu childfree, aku akan mendiduk anakku dengan baik". RT @noturdrugs: " sederhana, tp ga semua kucing bisa "	1

Figure 3: Bahasa Cyberbullying Dataset

This research produces Bahasa cyberbullying datasets that have a variety of writing procedures (syntax), the use of words with biased meanings, the delivery of words containing elements of pejoration (changing positive meaning into negative meaning), and non-standard language variations (the use of

foreign languages and slang languages). All of these characteristics can be found in Indonesian cyberbullying datasets. It will enhance and have a spread of data to determine the cyberbullying patterns that occur on social media (at least 10.5% of the data will contain cyberbullying patterns), and it will do this by drawing on the many different data collections that have been kept.

8. Exploring Document-term Using Large-Scale N-Gram for Capturing Cyberbullying Corpus Probability

The dataset on cyberbullying has certain inaccuracies, which results in changes in the quality of the data as well as the level of trust in the data. The reason that such flaws are present in the data may be linked to the fact that the data objects in question are notably dissimilar to or inconsistent with the typical data that is already there; in other words, there are outliers. In the process of building an Indonesian cyberbullying corpus, we performed feature extraction in order to investigate the search for document-terms.

The Bahasa cyberbullying corpus was prepared in phases till validation of data classification as cyberbullying and non-cyberbullying. The steps begin with Indonesian cyberbullying datasets and data separation. Data is separated by setting a testing size of 0.3 (70% training data, 30% validation/testing data). The Large-scale N-Gram approach is used to extract document-term features from every 1 word (n=1) to 5 words (n-5) successively. Each document line in the dataset is extracted many times.

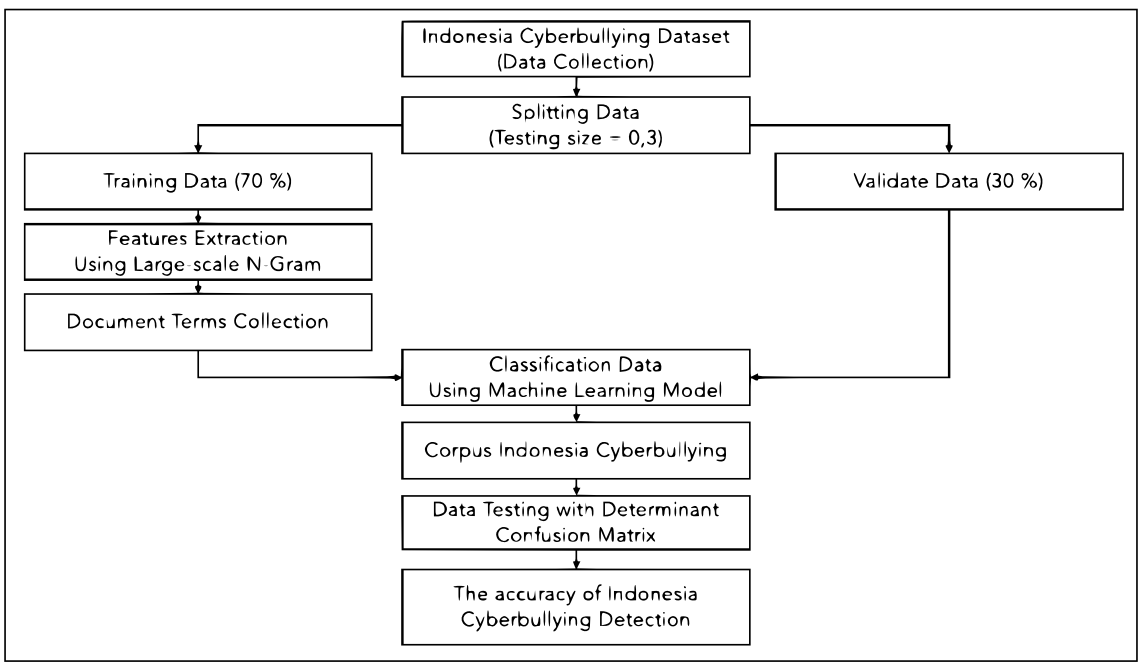


Figure 4: Workflow Phases for Creating and Analyzing The Bahasa Cyberbullying Corpus

The following step is carrying out the categorization procedure by comparing the document-term collection that was produced with the label of each individual document. The value of the Term Frequency-Inverse Document Frequency (TF-IDF) is used to classify and determine the labels for each of the terms that appear in the documents. The researchers also implemented smoothing strategies into the process, which helped them raise the probability value while calculating the TF-IDF value. The findings from this phase will be incorporated into the Bahasa Cyberbullying Corpus.

In addition to this, we do testing on the newly constructed corpus by using test data. The value of the confusion matrix is used to assess whether or not the Indonesian classification of cyberbullying is accurate. The computation of the confusion matrix is based on four components, which are as follows:

1. True Positive (TP): The model correctly predicted the positive class.
2. False Positive (FP): The model incorrectly predicted the positive class when it should have been negative.
3. True Negative (TN): The model correctly predicted the negative class.
4. False Negative (FN): The model incorrectly predicted the negative class when it should have been positive.

Typically, the Confusion Matrix is presented in table 2 format:

Table 2: Confussion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

To calculate accuracy using the confusion matrix, add the number of correct predictions (true positives and true negatives) and divide by the total number of predictions, using the equation 2:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{8.1}$$

Using the confusion matrix, the accuracy of the cyberbullying corpus can be calculated by testing against the documents in the final dataset. This study makes a good cyberbullying corpus that can be used for more research.

9. Machine Learning Model for Detecting Cyberbullying in Bahasa: an Evaluation

his part focuses on the implementation of machine learning techniques to accurately recognize instances of cyberbullying in the Indonesian context. This is achieved by utilizing a dataset and corpus that has been constructed specifically for this purpose. Following the preprocessing and feature extraction of the dataset, the classification phase is conducted using various classification algorithms, namely Support Vector Machine (SVM), Linear Regression, K-Nearest Neighbor (K-NN), and Random Forest. These algorithms employ distinct stages in the classification process. The last part will present the accuracy outcomes of each Large-Scale N-Gram investigation in comparison to machine learning techniques.

The N-Gram tested on a broad scale comprises four distinct patterns, specifically;
1. Large-scale N-Gram (N=1 to 2); the extraction process is done by extracting documents for every single word and every two consecutive words.
2. Large-scale N-Gram (N=1 to 3); the extraction process is done by extracting documents for every single word, every two consecutive words, and every three consecutive words.
3. Large-scale N-Gram (N=1 to 4); the extraction process is done by extracting documents for every one word, every two consecutive words, every three consecutive words, and every four consecutive words.
4. Large-scale N-Gram (N=1 to 5); the extraction process is done by extracting documents for every single word, every two consecutive words, and every three consecutive words, every four consecutive words, and every five consecutive words.

The accuracy testing findings of this research investigation are presented in Table 3. The findings indicate that effective and precise identification of Indonesian cyberbullying may be achieved (with the highest accuracy recorded at 99.92%) through the development of a Bahasa cyberbullying corpus and the utilization of a machine learning model for categorization purposes. The corpus on cyberbullying in Indonesia was created by employing the Large-Scale N-Gram technique to extract features. This approach enables the identification of document-terms commonly found in conversations or comments that involve cyberbullying. Moreover, it addresses challenges related to words with biased connotations, the use of foreign languages, and the presence of words that have undergone pejoration (a shift from positive to negative meaning) based on sentence syntax.

Table 3: Results of Testing the Accuracy of The Indonesian Cyberbullying Classification

	Sarcasm Dataset				Bahasa Cyberbullying Dataset (Source From Kaggle)			
Classifier	Large-scale N-Gram (1-2)	Large-scale N-Gram (1-3)	Large-scale N-Gram (1-4)	Large-scale N-Gram (1-5)	Large-scale N-Gram (1-2)	Large-scale N-Gram (1-3)	Large-scale N-Gram (1-4)	Large-scale N-Gram (1-5)
K-NN	82,94	83,03	83,19	83,19	79,49	79,49	79,49	79,49
Naïve Bayes	97,23	96,72	96,30	96,30	84,62	84,62	84,62	84,62
Support Vector Machine	99,92	99,92	99,92	99,92	75,39	75,39	75,39	75,39
Linear Regression	98,82	98,91	98,82	98,82	77,95	77,44	77,44	77,44
Random Forest	99,41	99,66	99,50	99,16	77,44	82,05	81,54	82,05
	Bahasa Cyberbullying Dataset (Crawling from Twitter)				Bahasa Cyberbullying Dataset (Crawling from Instagram)			
Classifier	Large-scale N-Gram (1-2)	Large-scale N-Gram (1-3)	Large-scale N-Gram (1-4)	Large-scale N-Gram (1-5)	Large-scale N-Gram (1-2)	Large-scale N-Gram (1-3)	Large-scale N-Gram (1-4)	Large-scale N-Gram (1-5)
K-NN	76,90	77,59	76,90	77,24	88,24	88,24	88,24	88,24
Naïve Bayes	78,28	76,55	73,79	71,72	83,01	84,31	84,31	84,31
Support Vector Machine	72,76	71,40	71,72	72,07	90,20	90,20	90,20	90,20
Linear Regression	75,17	75,52	75,86	74,83	90,85	90,20	90,85	90,20
Random Forest	75,52	76,55	76,55	78,28	93,46	92,81	94,12	91,50
	Bahasa Cyberbullying Dataset (Scrapping from Youtube)							
Classifier	Large-scale N-Gram (1-2)	Large-scale N-Gram (1-3)	Large-scale N-Gram (1-4)	Large-scale N-Gram (1-5)				
K-NN	91,60	91,58	91,56	91,57				
Naïve Bayes	88,59	89,15	89,36	89,47				
Support Vector Machine	86,98	86,84	87,09	87,23				
Linear Regression	89,77	89,95	89,96	89,97				
Random Forest	91,53	91,45	91,50	91,33				

The results of the experiments also show that the use of machine learning models is able to perform the classification process effectively. The results also show that each classification algorithm used in this study has a unique level of classification accuracy. This is due to the fact that each classification algorithm utilizes a unique series of stages and steps in order to complete the classification process.

When compared to other classification approaches, the machine learning models of Support Vector Machine (SVM) and Random Forest have a higher average accuracy. This is due to the fact that the SVM approach works by distinguishing between data in both a linear and non-linear fashion. The support vector machine (SVM) employs a transformation technique known as the kernel trick in situations when the data cannot be linearly separated. This brings the data to a higher dimension, which then makes linear separation possible. In addition, SVM can solve classification issues even when there is an imbalance in the data classes. The reason for this is that SVM gives priority to the largest margin in order to lessen the impact that the majority class has on the overall results. Although the Random Forest technique offers the benefit of merging the outcomes of a large number of decision trees that have been constructed individually, this approach's main advantage is that it is faster. Errors from each tree can compensate for each other thanks to the aggregation method, which also helps reduce the danger of the model being overfit to the data and improves its performance on fresh data. In addition to this, Random Forest is able to give information on the characteristics that have the greatest impact on the classification process. This is helpful in gaining an understanding of the function that qualities play in the process of prediction.

Using Large-scale N-Gram, we observed a specific capture pattern for document-terms. Large-scale N-Gram (word capture ordered from every one word to every four consecutive words) is shown to have

the highest accuracy in this test. The process of gathering terms to be used as a corpus may be carried out more correctly when document terms of up to four words are captured. However, if the capture of document-terms with growing high n values, this is not good and less accurate. Figure 5 presents a graphical representation of the precision of large-scale N-Gram classification tests.

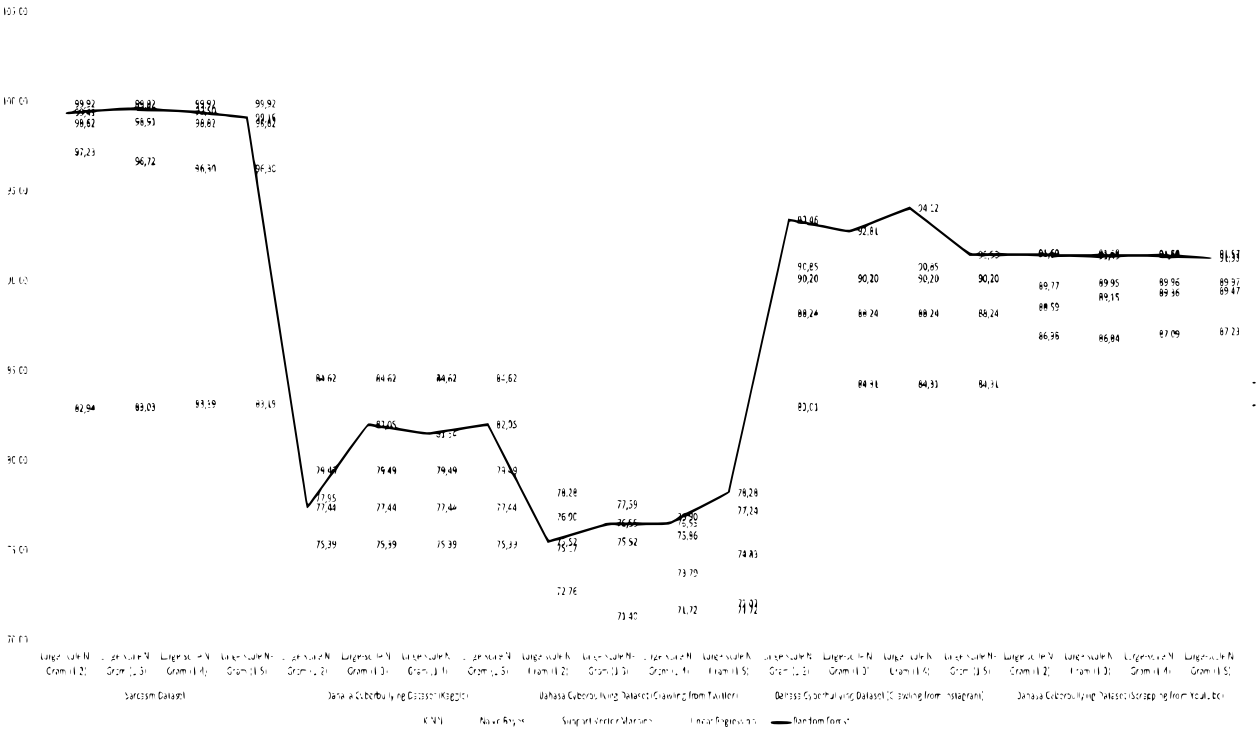


Figure 5: Graph of classification test accuracy with Large-scale N-Gram

10. Conclusion and Future Work

n this study, we explore the possibility that the method of word feature extraction might be used to identify instances of cyberbullying. A cyberbullying corpus may be trained using a variety of word permutations and combinations extracted from text using the Large-scale N-Gram approach. Machine learning in document classification can be used to deal with the linguistic diversity, biased word choice, and peyoration (the transformation of words with positive meanings into negative ones) that characterizes Indonesian cyberbullying conversations and comments.

Calculating the confusion matrix allows for an examination of the connection that has been established between the Bahasa cyberbullying corpus that has been constructed and the accuracy of cyberbullying document categorization. The findings that were obtained indicate that the Bahasa corpus that was extracted using the Large-scale N-Gram approach had the maximum accuracy of 99.92%, with the capture of terms that are made of a single word up to four words in a row.

Because of the process of training and testing on the Bahasa cyberbullying dataset with machine learning models, it is possible to construct an adaptive corpus and classification training process that follows the development of trends. This allows for the Indonesian cyberbullying corpus to be continually enhanced. Other algorithms are not as effective as the Support Vector Machine (SVM) and Random Forest algorithms when it comes to the process of document categorization. This is because these two algorithms extract features from documents more effectively.

Aside from the text arrangement (syntax) characteristic, more factors might be taken into consideration in further study. The purpose of this is to generate a data set consisting of different age groups, genders, and demographics in an effort to capture the determining patterns of cyberbullying on social

media. In order to generate detection accuracy that is both more accurate and superior, the utilization of additional methods or approaches is required as well.

References

1. Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 136, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

2. Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., & Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting TF–IDF. Computer Speech & Language, 56, 107–129. https://doi.org/10.1016/j.csl.2019.01.005

3. Baggini, J., & Fosl, P. S. (2010). The Philosophers (2nd ed.). Blackwell Publishing Ltd.

4. Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, 90, 101710. https://doi.org/10.1016/j.cose.2019.101710

5. Barlett, C. P. (2019). Chapter 2—Cyberbullying, Traditional Bullying, and Aggression: A Complicated Relationship. In C. P. Barlett (Ed.), Predicting Cyberbullying (pp. 11–16). Academic Press. https://doi.org/10.1016/B978-0-12-816653-6.00002-9

6. Chan, T. K. H., Cheung, C. M. K., & Lee, Z. W. Y. (2021). Cyberbullying on social networking sites: A literature review and future research directions. Information & Management, 58(2), 103411. https://doi.org/10.1016/j.im.2020.103411

7. Chawla, P., Hazarika, S., & Shen, H.-W. (2020). Token-wise sentiment decomposition for ConvNet: Visualizing a sentiment classifier. PacificVis 2020 Workshop on Visualization Meets AI, 4(2), 132–141. https://doi.org/10.1016/j.visinf.2020.04.006

8. Cohen-Shapira, N., & Rokach, L. (2021). Automatic selection of clustering algorithms using supervised graph embedding. Information Sciences, 577, 824–851. https://doi.org/10.1016/j.ins.2021.08.028

9. Genoud, A. P., Gao, Y., Williams, G. M., & Thomas, B. P. (2020). A comparison of supervised machine learning algorithms for mosquito identification from backscattered optical signals. Ecological Informatics, 58, 101090. https://doi.org/10.1016/j.ecoinf.2020.101090

10. I. Ting, W. S. Liou, D. Liberona, S. Wang, & G. M. Tarazona Bermudez. (2017). Towards the detection of cyberbullying based on social network mining techniques. 2017 International Conference on Behavioral, Economic, Socio-Cultural Computing (BESC), 1–2. https://doi.org/10.1109/BESC.2017.8256403

11. Imura, T., Toda, H., Iwamoto, Y., Inagawa, T., Imada, N., Tanaka, R., Inoue, Y., Araki, H., & Araki, O. (2021). Comparison of Supervised Machine Learning Algorithms for Classifying of Home Discharge Possibility in Convalescent Stroke Patients: A Secondary Analysis. Journal of Stroke and Cerebrovascular Diseases, 30(10), 106011. https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.106011

12. Li, Q. (2007). New bottle but old wine: A research of cyberbullying in schools. Computers in Human Behavior, 23(4), 1777–1791. https://doi.org/10.1016/j.chb.2005.10.005

13. Li, S., Pan, R., Luo, H., Liu, X., & Zhao, G. (2021). Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling. Knowledge-Based Systems, 218, 106827. https://doi.org/10.1016/j.knosys.2021.106827

14. López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., & Cacheda, F. (2021). Early detection of cyberbullying on social media networks. Future Generation Computer Systems, 118, 219–229. https://doi.org/10.1016/j.future.2021.01.006

15. Michael A, P., Sharon, R., Mats, H., & Tina, B. (2018). Post-Truth, Fake News. Springer International Publishing. https://doi.org/10.1007/978-981-10-8013-5

16. Noviantho, S. M. Isa, & L. Ashianti. (2017). Cyberbullying classification using text mining. 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), 241–246. https://doi.org/10.1109/ICICOS.2017.8276369

17. Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and Its Applications, 540, 123174. https://doi.org/10.1016/j.physa.2019.123174

18. Rajput, A. (2020). Chapter 3—Natural Language Processing, Sentiment Analysis, and Clinical Analytics. In M. D. Lytras & A. Sarirete (Eds.), Innovation in Health Informatics (pp. 79–97). Academic Press. https://doi.org/10.1016/B978-0-12-819043-2.00003-4

19. S. Salawu, Y. He, & J. Lumsden. (2020). Approaches to Automated Detection of Cyberbullying: A Survey. IEEE Transactions on Affective Computing, 11(1), Article 1. https://doi.org/10.1109/TAFFC.2017.2761757

20. Sharma, P., & Sharma, A. K. (2020). Experimental investigation of automated system for twitter sentiment analysis to predict the public emotions using machine learning algorithms. Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2020.09.351

21. Sheldon, P., Rauschnabel, P. A., & Honeycutt, J. M. (2019). Chapter 3—Cyberstalking and Bullying. In P. Sheldon, P. A. Rauschnabel, & J. M. Honeycutt (Eds.), The Dark Side of Social Media (pp. 43–58). Academic Press. https://doi.org/10.1016/B978-0-12-815917-0.00003-4

22. T. Mahlangu & C. Tu. (2019). Deep Learning Cyberbullying Detection Using Stacked Embbedings Approach. 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), 45–49. https://doi.org/10.1109/ISCMI47871.2019.9004292

23. T.K., B., Annavarapu, C. S. R., & Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. Computer Science Review, 40, 100395. https://doi.org/10.1016/j.cosrev.2021.100395

24. V. Banerjee, J. Telavane, P. Gaikwad, & P. Vartak. (2019). Detection of Cyberbullying Using Deep Neural Network. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 604–607. https://doi.org/10.1109/ICACCS.2019.8728378

25. W. M. Al-Rahmi, N. Yahaya, M. M. Alamri, N. A. Aljarboa, Y. B. Kamin, & F. A. Moafa. (2019). A Model of Factors Affecting Cyber Bullying Behaviors Among University Students. IEEE Access, 7, 2978–2985. https://doi.org/10.1109/ACCESS.2018.2881292

26. W. M. Al-Rahmi, N. Yahaya, M. M. Alamri, N. A. Aljarboa, Y. B. Kamin, & M. S. B. Saud. (2019). How Cyber Stalking and Cyber Bullying Affect Students' Open Learning. IEEE Access, 7, 20199–20210. https://doi.org/10.1109/ACCESS.2019.2891853

Yudi Setiawan, Study Program of Information System, Department of Engineering, University of Bengkulu, Bengkulu, Indonesia

E-mail address: ysetiawan@unib.ac.id