IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 10, 2022, accepted March 3, 2022, date of publication March 11, 2022, date of current version March 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158940

Contextual Text Analytics Framework for Citizen Report Classification: A Case Study Using the Indonesian Language

EVARISTUS DIDIK MADYATMADJA[1], **(Member, IEEE),**
BERNARDO NUGROHO YAHYA[2], **AND CRISTOFER WIJAYA**[1]
[1]Information Systems Department, School of Information Systems, Bina Nusantara University, West Jakarta 11530, Indonesia
[2]Industrial and Management Engineering Department, Hankuk University of Foreign Studies, Yongin, Gyeonggi 17035, South Korea

Corresponding authors: Evaristus Didik Madyatmadja (emadyatmadja@binus.edu) and Bernardo Nugroho Yahya (bernardo@hufs.ac.kr)

This study is supported by Initiative Project as a part of Bina Nusantara Professorship Grant 2022.

ABSTRACT Citizen science has emerged in many countries to contribute to the prompt resolution of individual field problems and has been shifted toward Information System (IS) research. In the domain of IS, a citizen report mechanism has been introduced in many local governments to understand regional problems based on the public participation. The rising of social media enforces many organizations including the local governments to utilize any information from the citizens including texts. Text mining has been utilized in various types of analyses such as sentiment analysis. However, it shows many challenges when it comes to the local context. The local context of words could cause various conflation errors that highly affect the learning task such as classification methods. This study aims to propose a context-based text processing and feed the proposed approach into a machine learning framework to classify the data of citizen reports-. The context-based text preprocessing utilized statistical- and semantic-based measurements to extract the local context and elaborate domain expertise to verify the misinterpretation for further text processing such as feature extractions. Subsequently, the n-gram language models together with the Term Frequency and Inverse Document Frequency schemes were performed to build the features. The result showed that the context-based text preprocessing improved the classification performance in majority classifiers in about 3% with the combinations of n-gram features.

INDEX TERMS Classification, e-government, public complaint, text mining.

I. INTRODUCTION

Citizen science has emerged as a way for the public to participate in scientific research. Sometimes, it is described as "public participation", "participatory monitoring", and participatory action research. While this emerging field was initially related to the nature, such as iNaturalist [1], the citizen science becomes important to bridge the society and the discipline of Information System (IS). A highlighted definition of citizen science in IS research was provided by [2] that mentioned "Citizen science in IS research is a partnership between IS researchers and people in their everyday lives".

Citizen reporting is one of the citizen science phenomena of Internet-based interaction between citizens and governments. By means of information and communication technology (ICT), the reporting Internet-based applications allow citizens to share information and/or knowledge that is relevant to the government services via web-based or mobile platforms [3]. In most of the applications, the platform has been used to report infrastructure issues at certain locations using geo-location technology [4] or public administration issues in the local government [5]. In the context of smart cities, citizen reporting is important due to the requirement of local governments to utilize a multitude of data sources to enhance or expand their services in "smart" ways.

For the citizen report, the use of public social media such as Twitter and Facebook helps to communicate with the citizens. Most of the time, the public channels were not found so efficient to handle citizen reports due to some issues such as private data and the difficulty to extract the proper data. Some

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

city governments developed a dedicated application for the citizen reporting for example My City Report in Japan, Fix My Street in some European countries such as UK, Canada, Norway, etc., Jakarta Qlue [6], and Tangerang Live [7]. The dedicated application could capture citizen voices based on their specific regions via built-in social media in Tangerang media using crowd-sourcing techniques. Since it is denoted as social media data, the use of text mining for scientific and social research e.g. to understand the social, economic, and historical processes is beneficial. Many works dealing with the citizen report emphasize the topic modeling and/or sentiment analysis from the public [8].

However, there are a few works on discriminatory analysis to handle problem categorization. To do the discriminatory analysis on problem categorization, there should be a proper text processing method. The text preprocessing such as stemming, stop-word removal is not sufficient to extract the proper feature when the culture and social contexts are considered for citizen reporting analysis. Local context word extraction causes more challenges in extracting and analyzing the citizen reports. Some works deal with the local such as Alexandria Contextual Text Analysis (ACTA), probabilistic latent semantic analysis (PLSA) [9], topic-specific networks [10], topic-based search engine [11], and temporal text mining [12]. In a specific language such as the Indonesian language, there are a lot of differences in terms of morphology. For example, changing or adding prefixes and suffixes in an Indonesian word can refer to meanings. The language diversity with more than 500 ethnics in such a country as Indonesia encounters problems to understand the context.

Reference [13] conducted a research in which they created a dataset of Bahasa Indonesia shortened terms that may be used to normalize any truncated words in Indonesian. Crowdsourcing was chosen as the approach for developing the dataset since only humans are capable of converting shortened words to their full form. 1063 of the 1170 sentences tested were correctly answered, while 107 were incorrectly answered. This research is about 90.85 percent accurate. Another study done by [14] focuses on developing a framework for preparing text mining applications by aggregating frequently used preprocessing methods. This system is organized around three primary preprocessing tasks: expansion, removal, and tokenization (ERT). The ERT reads the corpus and generates a list of tokens; these tokens are equipped to conduct all learning algorithms. The ERT architecture enables the rapid and accurate execution of all preprocessing procedures.

This study aims to propose a novel contextual text analytical framework to analyze citizen report (i.e. text) data. To extract the local context, we propose a contextual text processing approach which combines both statistical-based and semantic-based techniques. The statistical-based technique adopts the word cloud while the semantic-based technique utilizes the morphological analysis to understand the local words extracted from the word-cloud technique. The semantic-based technique utilizes two approaches: acronym2word retrieval and acronym-misinterpretation removal. In the end, a domain expert is involved in the analysis to improve the local context word extraction. To verify the proposed approach, we develop a machine learning framework and do various experiments to evaluate its performance. We perform the analysis with the inclusion of corpus in our context-based approach. At the stage before the classification, we elaborate the n-gram language models and do extensive experiments among several alternative methods to seek the best classification model. As a conceptual advance, we collaborate with partner local governments to enable public engagement under the smart city framework proposed by Tangerang, Indonesia. Hence, the contributions of this study are as follows:

- We propose a machine learning framework to classify citizen reports.
- We propose a context-based approach (called Con-TP) to minimize various conflation errors due to the existence of local languages by combining multiple techniques such as statistical-based and semantic-based analyses.
- We do extensive experiments on discriminatory analysis with various alternatives of n-gram language models to display the performance of the proposed context-based approach.

This paper is organized as follows. Section 2 addresses the related works. Section 3 explains the machine learning framework proposed in this study. Section 4 displays the report on using the proposed framework in the case study of Tangerang city. Finally, section 5 concludes this study.

II. RELATED WORK

A. PROBABILISTIC LATENT SEMANTIC ANALYSIS (PLSA)

Reference [15] proposed a novel unsupervised learning technique called Probabilistic Latent Semantic Analysis (PLSA), which is based on a statistical latent class model. The author claimed that this technique is more principled than conventional Latent Semantic Analysis since it is statistically sound. Additionally, the author empirically validated the potential benefits, generating significant performance boosts. Thus, Probabilistic Latent Semantic Analysis should be regarded as a novel unsupervised learning technique with a broad variety of applications in text learning and information retrieval.

B. TOPIC-SPESIFIC NETWORKS

Reference [10] demonstrates how to use text mining to extract topic-specific networks from blogs and how utilizing social network analysis on these topic-specific networks increases the efficacy and comprehension of blogger behaviors. Their solution is applying a topic detection text mining technique to their blog entries and then categorizes them according to the themes to which they most closely connect.

C. TOPIC-BASED SEARCH ENGINE

Within the context of topic-based search engines, we are researching applications for trend detection and analysis. In trend analysis, we are particularly interested in the temporal behavior of our system's automatically produced topic variables. We used a statistical topic model to analyze data from financial online newspapers. This model was then used to investigate the temporal behavior of the themes, which is the paper's primary contribution. We have shown that the model's subjects may reflect changes in current patterns and are congruent with popular perceptions of events. This method of using a search engine to get a better grasp of the world's temporal changes is both natural and convenient, since the need for search technology continues to grow [11].

D. TEMPORAL TEXT MINING

Temporal Text Mining (TTM) is the process of identifying temporal patterns in text data gathered across time. Due to the fact that the majority of text content has some type of time stamp, TTM has a wide variety of applications in a variety of disciplines, including summarizing news stories and exposing research patterns in scientific publications. In all of these instances, it would be advantageous to detect, extract, and summarize these evolutionary theme patterns automatically (ETP). Reference [12] examine the difficulty of identifying and summarizing ETPs in a text stream. They define the problem and present general probabilistic methods for resolving it by (1) identifying latent themes in text, which includes both interesting global themes and salient local themes for a given period; (2) establishing theme evolutionary relationships and constructing a theme evolution graph; and (3) modeling theme strength over time and analyzing theme life cycles.

III. RESEARCH METHODOLOGY

A. SOCIAL MEDIA ANALYTICS

In recent days, social media have been utilized to communicate among societal communities even they have been separated by a mile or so [16].The fast development of social media utilization has driven an expanding amassing of information which is increasingly known as Social Media Big Data [17]. The social media utilization creates new opportunities to analyze several aspects to gain insights into issues, patterns, persuasive actors, and other sorts of data. Social media have also been a critical driver for acquiring and spreading data in numerous sectors such a commerce, amusement, science, emergency administration, and legislative issues [16]. Social media are often used to express phenomena that happen in a certain location, so it can deliver exact information. The use of social media data to produce information that can be reviewed is called Social Media Analytics (SMA) [16]. SMA has been performed in various sectors such as tourism (i.e. urban smart tourism ecosystem) [18], hospitality (e.g. museum) [19], geographic sectors to analyze natural disaster management [5], healthcare in terms of exploring the adverse drug reaction of diabetes medicine [6], business sectors to improve companies' public relations [7], brand marketing [9], and national happiness [20].

Social media also enabled citizens to freely report their concerns via online approach to the respective company, related organization, or relevant government. There are several works on social media analytics related to public complaints. SMA had been implemented by extracting hotel ratings and reviews from Trip Advisor [21]. Another study also implemented SMA to understand customer experiences of the three largest drugstore chains in the United States with several frameworks such as quantitative analysis, text mining, and sentiment analysis [22]. This also serves as a roadmap for the Indonesian government in terms of developing computerized government services. Currently, the majority of phases in the development of e-government applications in Indonesia are focused on delivering websites and information application services [23].

Social media analytics receives the challenge in providing devices and systems to analyze the data of social media because each site of social media uses diverse platform volumes, complexity of the information, and unstructured data. People write words or sentences with errors. In order to let them write or search with legitimate grammar and structured sentences, the text mining approach is used. Text mining is like an intelligence system that extracts appropriate words or sentences from improper words and then changes those words into specific suggestions [24]. Text classification is a mandatory phase in the text mining beside the clustering and categorization phases that are useful for extracting knowledge as a starting point in text mining applications. Text classification techniques can be divided into statistical and machine learning (ML) approaches. The statistical techniques purely fulfill the manual hypothesis, so they require little algorithm.

However, the ML techniques were specifically created for automation. There are several categories of algorithms based on the learning criteria i.e. the supervised, unsupervised, and semi-supervised categories. The supervised classification algorithms are further divided into two categories based on the supremacy of parameters in the data namely parametric (e.g. logistic regression and Naïve Bayes) and non-parametric (e.g. Support Vector Machine (SVM), Decision Tree, Rule Induction, K-nn, and Neural Networks). In this study, we used several types of supervised algorithms that are k-Nearest Neighbor (kNN), Random Forest, Support Vector Machine (SVM), Naive Bayes, and AdaBoost to compare the types of algorithms getting the best text classification results based on their percentage [25].

B. APPLICATION-BASED PUBLIC COMPLAINTS

Social media are ideal means of information to measure public opinion on policy from a political sector perspective. A good citizen complaint and feedback mechanism is an important aspect of providing quality public services because it can make the government more effective and responsive to the needs and demands of the public both in the policy and political realms. Public complaints are an opportunity for the government to improve themselves. Nowadays, public complaints are often submitted through various communication platforms such as social media managed by the institution [26]. The use of social media by citizens opens a big opportunity for public complaint administration bodies. It is proven by several studies of social media-based public complaints. Marko outlined that the use of social media has a positive relationship with social capital, civic engagement, and political participation; while Arturo analyzed the use of various social media i.e. Facebook and Twitter in terms of participating in local government issues [27], [28]. Raimundo described that social media are effective tools for society, able to set political agenda and influence political discourse [29]. Social media are said to have the potential for increasing political participation and discussions among citizens. The use of social media allows the government to know the opinions of the citizens about an issue so that the government can make the appropriate policies. This process makes the governance process more effective and efficient [30]. Good handling of public complaints can increase the transparency and accountability of the government. A transparent organization generates trust, and in the end encourages citizens to be more involved and participate in the policy-making process [26].

The information in social media can be filtered by crowd-sourcing [30]. Several studies using crowd-sourcing have been found improving government performance by facilitating relationships between public professionals and citizens such as Challenge.gov and Next Stop Design which adopt crowd-sourcing to solve specific public problems in the United States [31]. In Indonesia, government sectors have increased the use of social media as their communication channels. For example, Jakarta Provincial Government used Qlue and CROP as an information system in social media platforms to receive the public complaints and improve public participation in the development of Jakarta City [6]. There are several text minings on public complaints in social media such as the case study on Indonesian-language tweets related to obesity [32], the text mining case study related to the 2017 round 2 of the DKI Jakarta regional elections [33], and the text mining implementation based on Twitter data to analyze information regarding corona virus in Indonesia [34] in which the whole study uses the Term Frequencies method. In the case study on Indonesian-language tweets related to obesity and based on the level of occurrence of words and visualization, netizens tend to use words that are relevant to the Indonesian language (''gemuk'' or ''gendut''). The tweet originally used Indonesian language preferences and is more commonly used than the word ''obesity'' which is a word from a foreign language uptake [32].

In this study, public complaints on the Indonesian government social media will be analyzed. For the analysis, we will use Sastrawi which is one of the most popular corpus for the mining in the Indonesian language [32], [35]. This corpus has a standard in stop-words and is unable to stem some local or contextual words. According to Arisanti's research on the Indonesian language use on Facebook, acronyms and/or slangs are used more frequently on social media (e.g. Facebook) than standard vocabulary or abbreviations in Indonesian [36]. The acronyms and/or slangs often hinder the performance analysis since those words are unrecognizable by the corpus [37]. To enhance the analysis performance, this study develops an approach namely context-based text processing (Con-TP) to extend the existing corpus (i.e. corpus-based text processing (Cor-TP) to fit with the contextual words.

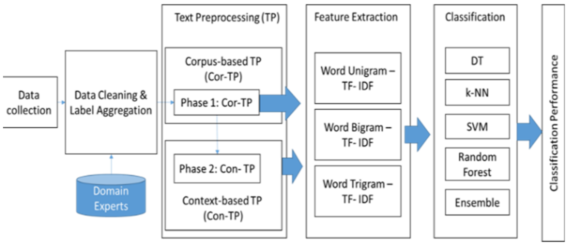


FIGURE 1. Machine learning framework.

IV. THE PROPOSED FRAMEWORK

This section aims to show a machine learning framework to analyze the data of public complaints. The proposed framework is as depicted in Fig. 1. The framework starts with data collection and data cleaning. In term of label aggregation, it was an elaborated task with domain experts. The cleaned data feeds into text preprocessing phase before we perform feature extraction. In the final stage, classification tasks with various classifiers are performed. The details of each of the phase will be explained in the later section. In particular, the detailed explanation will focus on our contribution which is the Context-based TP (Con-TP).

A. DATA COLLECTION

This study utilizes the citizen reports from two social media sources of the Government of Tangerang, Indonesia. The first data were collected from an application named LAKSA Tangerang LIVE which contains the complaint data from the citizens of Tangerang City. The second data were obtained from the Tangerang Government social media through comments or inbox features that contained public complaints. It should be noted that we utilize the data set from the LAKSA Tangerang Live application since it contains the categories of the citizens' problems. Meanwhile, the data set obtained from social media has no labels or categories.

B. DATA CLEANING AND LABEL AGGREGATION

The collected data contain inaccuracies, missing data, code errors, and other problems. Data cleaning plays a major role in improving the data quality prior to doing the analysis [36]. The data with null values are omitted. For the training purpose, we utilize the data from the LAKSA application which provides the labels of the citizen problems. To annotate the citizen reports, we perform label annotation and label aggregation to each of the citizen reports based on the domain experts. In this case, we conduct focus group discussions with some government members to determine the proper labels for the unstructured data.

C. TEXT PREPROCESSING

The collected data contain inaccuracies, missing data, code errors, and other problems. Data cleaning plays a major role in improving the data quality prior to doing the analysis [36]. The data with null values are omitted. For the training purpose, we utilize the data from the LAKSA application which provides the labels of the citizen problems. To annotate the citizen reports, we perform label annotation and label aggregation to each of the citizen reports based on the domain experts. In this case, we conduct focus group discussions with some government members to determine the proper labels for the unstructured data.

The text preprocessing step is a phase to perform various techniques to prepare the data prior to the feature extraction phase [38]. Generally, the step consists of approaches such as stemming and stop-word removal, is based on the corpus, and called the Corpus-based Text Processing (Cor-TP). To handle the local context of words, this study proposes a phase called the Context-based Text Preprocessing (Con-TP). The two phases are consecutive phases with the former being the Cor-TP and the later the Con-TP.

1) PHASE 1: CORPUS-BASED TEXT PROCESSING (COR-TP)

Phase 1 refers to the Corpus-based Text Processing (Cor-TP) and utilizes an Indonesian corpus named Sastrawi. Sastrawi is a library to stem and reflect the Indonesian language (words) from the base forms [39]. This corpus provides alternative words in a large corpus to seek proper syntactic patterns that occur with a seed list of opinion words [40]. This phase consists of five general stages: case folding, filtering, tokenizing, stemming, and stop-word elimination.

Case Folding is the process of changing all letters in the text into similar letters [38]. In this case, authors decide to change all letters to lowercase letters. Filtering process is performed by removing all non-alphabetic characters [38]. The characters that are all deleted are dots, commas, colons, and others. Tokenizing is the process of converting sentences into tokens by separating each word into a token [38]. The example of tokenizing is the sentence ''I want to complain'' changed to ['I', 'want', 'to', 'complain']. Stemming is a process in the text preprocessing that aims to decompose the various forms of a word into its base form by removing the affixes. In Indonesian languages there are prefixes, suffixes, infixes and confixes that will make a basic word change into various forms, and the search of basic words will be increasingly difficult, so the stemming stage is very necessary before the text mining process.

In this study, the stemming stage is carried out using the Sastrawi Indonesian stemmer library which is the newest stemmer [41]. Stop-word elimination is the stage of eliminating meaningless and unimportant words in a sentence. The word elimination aims to facilitate a program in understanding the structure and meaning of a sentence [37]. In this study, the elimination stage is carried out using the Sastrawi Indonesian stop-word remover library.

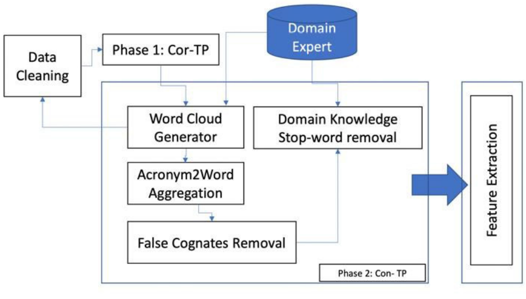


FIGURE 2. Phase 2: Con-TP.

2) PHASE 2: CONTEXT-BASED TEXT PREPROCESSING (CON-TP)

Phase 2 refers to the Context-based Text Preprocessing (Con-TP) and utilizes the elaboration between the statistical- and semantic-based approaches. The proposed Con-TP aims to improve the performance of Phase 1 (Cor-TP) and reduce the ambiguity in the use of slangs/abbreviations/acronyms. The acronyms and slangs are diverse in accordance to the geographical areas although the citizens share the same language [42]. For example, some of the Americans mentioned a beverage as ''coke'',while those from other regions mentioned it as ''tonic'' [43]. It is the same as Indonesian where there are about 500 different ethnics [44].

This phase consists of four stages as shown on Figure 2: word-cloud generator with frequency, acronym2word aggregation, acronym misinterpretation removal, and domain knowledge stop-word removal. The details of each step are explained in the later section.

3) WORD-CLOUD GENERATOR WITH FREQUENCY

A word cloud is a kind of list for visualizing language or text data [45]. The word-cloud prototype was originally used as a graphical map to show the relative sizes of regions in terms of relative font sizes [46]. Word cloud has a prominent result to show some important messages such as enrichment analysis [47], hospital admission diagnosis [48], and customer review [49]. There are two types of word cloud: visual features of tag cloud and different layouts of tag cloud comparison [50]. In this study, the type of word cloud used is the visual features of tag cloud by considering the frequency. In the frequency type, the font size represents the number of keywords that appear in the data. The larger the word size is, the more frequently the words appear so that by looking at the size, a decision can be made about whether the frequently appearing data is used, deleted, changed, etc. [51].

This stage is to generate word cloud from the data cleaning phase. We proceed the repetition process of the phase 1 and word cloud generator before entering the next stage. The result of the word cloud is further checked through the phase 1 (Cor-TP) and fed into the word cloud generator again for re-check. The involvement of domain experts is necessary to improve the result of phase 2 in particular the stage of the word cloud generator.

The word cloud from the data cleaning stage is constructed in this stage. Before proceeding to the next phase, we repeat the first phase procedure with the word cloud generator. The word cloud outcome is double-checked in phase 1 (Cor-TP) before being fed back into the word cloud generator. To improve the results of step 2 particularly the stage of the word cloud generator, domain experts must be included.

4) ACRONYM2WORD AGGREGATION

Acronyms are abbreviations that consist of the first letter or first few letters of words in a phrase commonly used in web searches, electronic communications, and social media [52]. The use of acronyms is very common due to the practical typing on mobile devices. There are two types of acronyms, (1) acronyms in the form of a combination of the initial letters of a word series (for example BPM, ERP), (2) acronyms in the form of a combination of syllables or a combination of letters and syllables from a word series (for example Radar (radio detection and ranging)) [53].

This stage follows the result of the word cloud. Most of the words extracted by the word cloud are acronyms due to the aforementioned reasons. Acronyms evolve dynamically from day to day and make themselves ambiguous because the same acronym has many different meanings. The acronym ''BPM'' could have different meanings without further explanation. ''BPM'' can be interpreted as Business Process Management, while the other meaning is ''Beat Per Minute''. In Indonesian, there are many similar cases. ''KKN'' can be interpreted as ''Korupsi, Kolusi, dan Nepotisme'' which means corruption, collusion, and nepotism. In addition, it can also be interpreted as ''kuliah kerja nyata'' which isa form of community service activities by students with a cross-scientific and sectoral approach at certain times and regions in Indonesia. Another example is the ambiguous acronym ''KTP'' short for ''Kartu Tanda Penduduk'' (Identity Card) in the Indonesian language. When the user types ''Kartu Tanda Penduduk'', there is a high possibility of typos such as ''Krtu Tanda Penduduk'' that may affect the performance such as classification and text recognition. In addition, the word ''Krtu'' will initially be considered differently, while the ''Tanda Penduduk'' is still recognized from ''Kartu Tanda Penduduk'' although the points are not maximal.

5) FALSE COGNATE REMOVAL

Cognates refer to words extracted from the same [54]. However, not all cognates have the same meaning. For example, the words ''camat'' and ''lurah'' are the results of stemming from the Sastrawi library, each of which can be interpreted as 2 words that have different meanings. The Stemmer Sastrawi library converts ''kecamatan'' and ''kelurahan'' into ''camat'' and ''lurah''. This is ambiguous because ''kecamatan'' and ''kelurahan'' are defined as words that indicate areas, while ''camat'' and ''lurah'' are defined as words that indicate persons. Therefore, the words ''kecamatan'' and ''kelurahan'' are manually changed to ''kecamatan'' and ''kelurahan'' as the aggregation of writing addresses or regions.

6) CONSTRUCTION OF A NEW STOP-WORD LIST FOR REMOVAL

Stop-words are those that appear frequently in the data, but are meaningless and unimportant for analytical value in the text mining. This step is made to eliminate words that should be eliminated by Sastrawi in phase 1. The use of abbreviations affects the stop-word removal process where words that should be removed are not detected [55]. For example, the word ''yg'' which refers to the abbreviation of ''yang''. The Sastrawi stop-word eliminates useless words in common. Common words in text documents such as prepositions, nouns, etc. which do not give any meaning to the document [56]. For example, the words ''and'', ''you'', ''when'', ''what'' are deleted because they are not considered as important words.

Sastrawi does not eliminate words that specifically lead to a subject. In this study, words related to public complaint are not eliminated. In public complaint data, not all words related to public complaint are needed. Location indicator words are not needed because this study classifies complaints only by referring to Tangerang City (sub-district). Some call words for Tangerang City government officials such as ''Sir'', ''Mr'', ''Headman'' etc. are not so important that they need to be eliminated.

Word sense disambiguation (WSD) is a fundamental natural language processing task [57]. Ambiguity exists when there are many alternatives of linguistic structures that can be composed for an input language [58]. For example, the word 'pagi' in the informal Indonesian language interprets 2 meaning: time interpretation and greetings.

7) FEATURE EXTRACTION

Feature extraction is an important step to perform the machine learning approach. In the domain of text mining, the extracted text is transformed into variables on which the machine learning, such as classification, can be performed. The n-gram model is one of the approaches to create bag-of-words (BoW) to represent the sequence of the words regardless of the grammar and word order [59]. The n-gram calculates the probability of the extracted features based on the probabilistic function [60]. Unigram represents the n-gram of size 1, Bigram represents the n-gram of size 2, and Trigram represents the n-gram of size 3. A higher n-gram represents four-gram, five-gram, and so on [61]. This study uses n-gram with the maximum range of $n = 3$ followed by the combinations of TF-IDF.

$$\boldsymbol{Unigram}: \boldsymbol{P}(\boldsymbol{w}_1, \ldots \boldsymbol{w}_n) = \prod\nolimits_i \boldsymbol{P}(\boldsymbol{w}_i) \tag{1}$$

$$\boldsymbol{Bigram}: \boldsymbol{P}(\boldsymbol{w}_i|\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i-1}) \approx \boldsymbol{P}(\boldsymbol{w}_i|\boldsymbol{w}_{i-1}) \tag{2}$$

P : probability
w : word
Extend to trigram, 4-gram, etc.

Since the general n-gram takes into account the number of occurrences, it would hinder the classifier when there are almost similar words. One of the standard tools in document classification is Term Frequency Inverse Document Frequency (TF-IDF). TF-IDF is a statistical measure to assign a weighting scheme to a word by finding how significant a word is to a document by giving weightage to a word in the document [62]. The TF-IDF method usually calculates the weight of each word to find the priority of the word in the corpus. The score will be higher when there are words that appear frequently [63].

$$\boldsymbol{tf} = \boldsymbol{f_d}(\boldsymbol{w}): \textbf{frequency of } \boldsymbol{w} \textbf{ in document} \tag{3}$$

where *tf* denotes the term frequency and $f_d(w)$ is the frequency of words in a document. The commonness of a word can be diminished through the weight of terms that occur very frequently in the document set and increased the weight of terms that occur rarely using inverse document frequency (*idf*) which is denoted in Equation (4).

$$\boldsymbol{idf}(\boldsymbol{w}, \boldsymbol{D}) = \boldsymbol{log}\frac{1 + |\boldsymbol{D}|}{1 + \boldsymbol{df}(\boldsymbol{d}, \boldsymbol{w})} \tag{4}$$

where *idf* denotes the inverse document frequency, *w* refers to word, *d* refers to a particular document, *D* is a set of documents, and $|D|$ is the number of documents in the set. The recommended words can be calculated using term-frequency-inverse document frequency (*tfidf*) based on the Equation (5).

$$\boldsymbol{tfidf}(\boldsymbol{w}, \boldsymbol{d}, \boldsymbol{D}) = \boldsymbol{tf}(\boldsymbol{w}, \boldsymbol{d}) \times \boldsymbol{idf}(\boldsymbol{w}, \boldsymbol{D}) \tag{5}$$

The *tfidf* is obtained by multiplying the *tf* and *idf*. In this study, the document refers to a particular complaint from a subject or a person. The TF-IDF result is then normalized with L2 norm (Equation (6) to understand the variance of the word variation from the complaint data.

The TF-IDF result is then normalized with L2 norm to understand the variance of the word variation from the complaint data.

$$v_{norm} = \frac{\hat{v}}{\|\hat{v}\|_2} = \frac{\hat{v}}{\left(\sum_{i=1}^{n} |\hat{v}|^p\right)^{\frac{1}{p}}} \tag{6}$$

v_{norm} : normalization vector
$\hat{v}$: unit vector
p : Euclidean norm

D. CLASSIFICATION

The classification in the data mining is useful for classifying each item in a data set into a small data set in a class or group that has been predetermined [64]. Text classification is one part of the classification that aims to classify text documents or natural language into a set of predefined categories and is often used in the fields of information retrieval, search engine, question answering system, public opinion analysis, and emotional analysis [65]. Every technique in the classification is used to support a learning algorithm to find the best fit relationship between the attribute set and class label of the input data [64]. This study used several types of classification technique algorithms namely k-Nearest Neighbor (kNN), Random Forest, Support Vector Machine' (SVM), Naive Bayes, and AdaBoost.

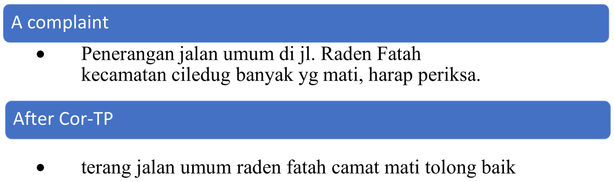
k-Nearest Neighbor is an instance-based classification algorithm that converts categorical attributes to numerics [55]. The kNN working principle is by classifying data based on the distance from each neighbor or the closest class [54]. Random forest is an algorithm based on ensemble learning that joins different types of algorithms or the same algorithm multiple times to get a more powerful prediction model (i.e. multiple decision trees) [66]. The more trees there are in the forest, the more robust and higher the accuracy of the prediction will be obtained [63]. Support Vector Machine (SVM) is a supervised learning algorithm that is based on a statistical theory. Support vector machines (SVMs) are a promising machine learning technique that has demonstrated excellent performance in the majority of prediction tasks [67]. The experiment done by [68] was combining SVM with Resnet50 demonstrated the greatest classification rate in experiments. This algorithm builds models that assign new data sets into a single category or makes it a non-probabilistic binary linear classifier [63]. Kernel functions offer the ability of computing data points in higher Linear, Sigmoid, and Polynomial [69]. Naïve Bayes provides classification parameters and attributes to label the occurrences. This algorithm is suitable for moderate or large training data sets [70]. AdaBoost is an important ensemble learning method that combines several weak classifiers and integrates them to be one strong classifier [58].

V. RESULT AND DISCUSSION

This section aims to display the analysis result and discuss the analysis from the text classification. The analysis result has been performed using a dataset from one of the city government in Indonesia. Afterward, we discuss the analysis result in accordance to the performance of the text classification results.

A. DATASET INFORMATION

In this study, we perform the analysis using a data set from the LAKSA application in Tangerang City. The citizen report data from the LAKSA application contain 9865 events and have 320 categories. The category refers to the label for the classification. According to the domain experts, the data sets can be aggregated and classified into four (4) maincategories: disaster, infrastructure, social, and nation-related affairs with the portions of 4%, 38%, 53%, and 4%, respectively. In the data cleaning process, we eliminate data without category and turn them into 6321 events.



A complaint

- Penerangan jalan umum di jl. Raden Fatah kecamatan ciledug banyak yg mati, harap periksa.

After Cor-TP

- terang jalan umum raden fatah camat mati tolong baik

FIGURE 3. The text preprocessing phase 1 (Cor-TP).

TABLE 1. Acronym2Word aggregation.

The acronyms and abbreviated words	Change into	Meaning
ektp, ktp	kartu tanda penduduk	ID Card
pju	penerangan jalan umum	Public Street Lighting
pkl	pedagang kaki lima	Street Vendor
nik	nomor induk kewarganegaraan	National Identification Number
kk	kartu keluarga	Family Registers
pbb	pajak bumi bangunan	Property Taxes
jln	jalan	Street

B. TEXT PREPROCESSING PHASE

The implementation of the Cor-TP with an example of public complaint "Penerangan jalan umum di Jl. Raden Fatah Kecamatan Ciledug banyak yg mati, harap periksa." can be seen in Figure 3. The results obtained in this stage are described in Figure 2.

The Cor-TP modified and removed the words of the complaint data as many as 34,453 words from the 172,095 total words in the complaint data or 26% of the total complaint data. In Figure 3, several words do not seem to have much impact on the classification performance such as "yg", "ada", "mohon", "banyak", etc. The Cor-TP changed several words such as "kecamatan" and "kelurahan" to "camat" and "lurah" which produce false cognates. Furthermore, the word cloud shows words "ektp" and "ktp" which have the same meaning, and the word "kartu" which could indicate "kartu tanda penduduk" (ktp).

E-ktp and KTP have the same meaning as the "Kartu Tanda Penduduk" (identity card). This can happen because many people use abbreviated words in the complaint data. When classified, words that have different forms even though they have the same meaning will still be considered different, so the unifying word process is needed to make no differences in several words that have the same meaning. In this study, the aggregation Cor-TP was re-applied to the acronyms and abbreviated words to make their classification more accurate. The results can be seen in Table 1.

TABLE 2. The stemming result for the word that has same root.

Words before stemming		Words after stemming	
Words	Meaning	Words	Meaning
kecamatan	address or region	camat	head of the district
kelurahan	address or region	lurah	head of the village

TABLE 3. The words that are not included in stop words in Sastrawi.

Words	Reasons
Sore, Malam	Public street lights are often associated with night (malam) conditions
Jalan	Jalan (roads) are often used associated with road conditions such as potholes, traffic jams

It turns out that the stemming process generates false cognates. The Sastrawi stemmer library, for example, will transform the terms "kecamatan" and "kelurahan" to "camat" and "lurah" The prefix "ke-" and suffix "-an" form the confix "kecamatan." The confix "ke—an" is used to classify words as nouns [71]. As a result, "kecamatan" and "camat" are considered nouns. Despite the fact that both terms are nouns, the derivation of "kecamatan" from "camat" is indicated by a change in the reference [60]. This appears in the term "kelurahan" as well.

This is problematic since the nouns "kecamatan" and "kelurahan" are defined as location nouns, whereas the nouns "camat" and "lurah" are classified as human nouns (shown in Table 2). This shift in reference implies that the lexical identities of the two lingual forms are distinct [68]. As a result, stemming was used once more to manually replace the phrases "kecamatan" and "kelurahan" with "kec" and "kel" as the aggregates of addresses or regions.

The next step is to re-implement the stop-word removal procedure by including manual stop-words that remove stop-words that are unused by the Sastrawi library. The name of the area/city, the word that shows the address/location indication, the call word, the greeting word, the pronoun, the question word, and other terms that are considered useless are among the manual stop-words that are added to be erased.

TABLE 4. The result of the classification performance by precision.

Framework		Uni	Bi	Tri	Uni+ Bi	Bi + Tri	Uni + Bi + Tri
Cor-TP (percent)	**K-NN**	62,00	37,93	41,03	61,86	40,29	62,58
	Naive Bayes	47,92	52,27	61,92	52,94	52,60	54,30
	SVM - Polynomial	74,29	61,70	50,87	73,60	60,34	73,48
	Random Forest	62,21	56,14	50,58	62,50	55,71	61,96
	Adaboost–RF	62,83	56,41	50,25	63,20	56,22	62,69
Cor-TP + Con-TP (Proposed Framework)	**K-NN**	63,34	41,98	46,57	62,26	45,28	63,44
	Naive Bayes	49,72	52,62	46,53	54,32	54,62	54,52
	SVM - Polynomial	77,01	61,09	50,74	76,55	60,86	75,94
	Random Forest	66,19	56,85	50,64	67,25	56,49	65,92
	Adaboost – RF	67,13	57,85	50,29	66,73	56,45	66,93

The classification performance by Precision for the proposed framework shows more highlighted than Cor-TP, which means the proposed framework indicates a better result.

Table 3. shows the words that are not included in stop words in Sastrawi

This study uses TF-IDF as a feature extraction which combines with N-gram. Indonesian words often use 3-letter acronyms (e.g. "kartu tanda penduduk", "surat izin mengemudi"). Therefore, the maximum range of N-gram is three which produces 6 features namely:

- unigram
- bigram
- trigram
- unigram + bigram
- bigram + trigram
- unigram + bigram + trigram

C. CLASSIFICATION RESULT AND PERFORMANCE

The classification performance is measured using several algorithms. F1-score is treated as the comparison value. We compare the performance between the Cor-TP and the Cor-TP + Con-TP. Before the classification, we split the data into training and testing data with the 80:20 ratio. We performed the empirical results with different numbers of parameters for each algorithm and return the best F1 score results among the parameter values.

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{7}$$

TP : True Positive
FP : False Positive

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{8}$$

FN : False Negative

$$\boldsymbol{F1 - Score} = \frac{\boldsymbol{2 + precision * recall}}{\boldsymbol{precision + recall}}$$

Majority voting is a method that uses multiple machine learning models that are integrated to produce output predictions based on the majority vote from all the models [72]. To achieve the best results, this technique requires at least a tree machine learning model [73]. Each base classifier outputs simply the label for the majority [74]. The best appropriate n-gram model for the categorization will be determined by the majority voting.

Figure 3 depicts a voting procedure involving a machine learning model with three computers in this case. First, the data from the training set will be used to feed into each learning model. Furthermore, the data will be analyzed using machine learning to generate predictions for each machine learning model voting outcomes. Third, once each classification model projected results have been obtained, the best appropriate n-gram model will be determined by a vote based on the outcomes of the most votes.

Our analysis shows that, somehow, the classification F1 score with two phases (e.g. Cor-TP and Con-TP) slightly outperforms the scenario with only the Cor-TP. The result shows that the combination of the Proposed Framework with three n-grams gives the best F1 score result among the respective combinations. Table 4. shows the result of the classification performance by precision. Table 5. shows the result of the classification performance by recall. Table 6. shows the result of the classification performance F1 score.

The results of this study suggest that the classification F1 score with two phases (e.g. Cor-TP and Con-TP) outperforms the scenario with only the Cor-TP by a small margin. The study indicates that the Proposed Framework in combination with three n-grams produces the best F1 score among the various combinations.

Unit of each algorithm and n-gram. The highest F1 score is found in the combination of the SVM with the polynomial

TABLE 5. The result of the classification performance by recall.

Framework		Uni	Bi	Tri	Uni+ Bi	Bi + Tri	Uni + Bi + Tri
Cor-TP (percent)	K-NN	91,67	72,59	70,85	92,67	74,90	92,94
	Naive Bayes	91,06	90,85	83,30	91,71	73,81	91,87
	SVM - Polynomial	90,60	82,44	70,85	89,24	80,65	87,36
	Random Forest	88,78	83,12	73,55	89,88	81,86	88,37
	Adaboost–RF	91,33	87,87	7,63	86,94	79,13	87,55
Cor-TP + Con-TP (Proposed Framework)	K-NN	89,20	80,56	80,90	92,93	79,17	93,32
	Naive Bayes	91,62	91,19	88,18	91,88	90,97	79,27
	SVM - Polynomial	89,05	83,12	68,79	88,06	83,28	86,73
	Random Forest	87,51	78,10	72,56	88,53	80,02	88,63
	Adaboost – RF	88,96	78,48	74,10	89,19	78,28	89,11

The classification performance by Recall for the proposed framework shows more highlighted than Cor-TP, which means the proposed framework indicates a better result.

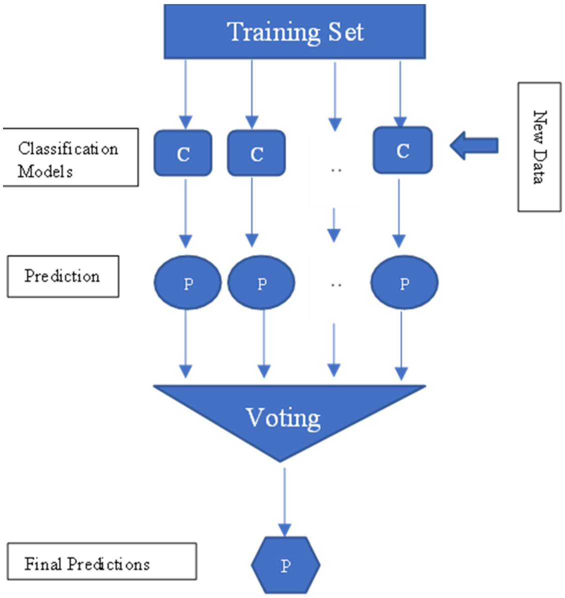


FIGURE 4. Classification process.

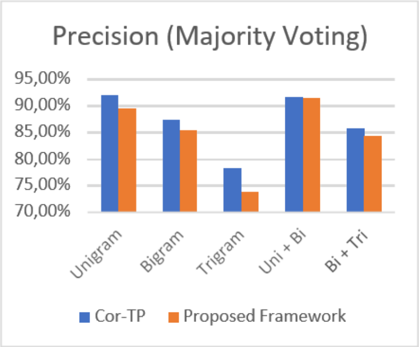


FIGURE 5. The result of the classification performance by precision using majority voting.

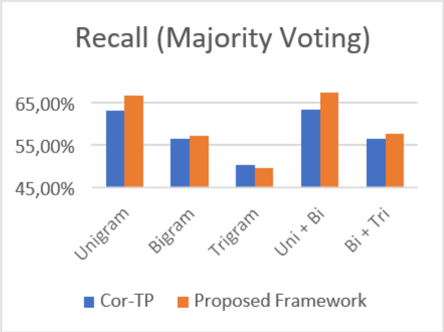


FIGURE 6. The result of the classification performance by recall using majority voting.

kernel algorithm and unigram in the proposed framework which has 81,83% F1 score.

Figure 5, 6, 7 shows that in majority voting the combination of unigram, bigram, and trigram has the highest F1 score compared to the others which is 74,35%. In a single feature, a unigram has the highest F1 score compared to other single features: bigram and trigram. In the proposed framework, a decrease occurs when unigram data sets changed to bigram data sets. It also occurs when the data sets changed to trigram data sets. The total decrease from unigram to trigram is 15%. This also occurs in the Cor-TP with the total decrease from unigram to trigram is 14%. Oppositely, if unigram data sets are combined with bigram data sets, there is an increase of the F1 score compared to a single feature in n-gram data sets. However, a decrease occurs when the data sets from the combination of unigram and bigram changed into the combination of bigram and trigram.

TABLE 6. The result of the classification performance by F1-Score.

Framework		Uni	Bi	Tri	Uni+ Bi	Bi + Tri	Uni + Bi + Tri
Cor-TP (percent)	K-NN	68,70	32,29	40,24	68,77	36,20	69,60
	Naive Bayes	48,39	54,96	49,33	55,42	55,45	57,01
	SVM - Polynomial	80,22	67,76	55,87	79,39	66,26	78,76
	Random Forest	68,72	62,26	55,84	69,31	61,78	68,45
	Adaboost–RF	69,75	62,89	54,90	69,25	61,91	69,10
Cor-TP + Con-TP (Proposed Framework)	K-NN	70,06	37,89	46,84	69,56	43,19	70,78
	Naive Bayes	50,91	55,59	50,09	57,42	58,90	57,67
	SVM - Polynomial	81,83	67,37	55,01	81,18	67,29	80,34
	Random Forest	72,45	61,93	55,56	73,64	62,31	72,32
	Adaboost – RF	73,69	63,19	55,45	73,22	61,93	73,56

Theclassification performance by F1-Score for the proposed framework shows more highlighted than Cor-TP, which means the proposed framework indicates a better result.

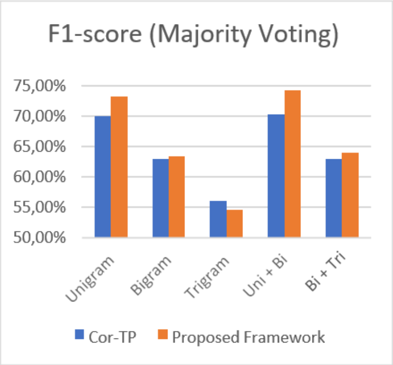


FIGURE 7. The result of the classification performance by F1-Score using majority voting.

SVM with a polynomial kernel using unigram is implemented on uncategorized data from social media. The uncategorized data from social media are classified as 2% of disaster, 27% of infrastructure, 61% of social, and 10% of nation-related affairs. The output shows many of the data are categorized in social class. It might occur because the training data have an imbalanced amount of data in each category. It can also lead to unexpected results.

VI. CONCLUSION

This study aims to propose a context-based text preprocessing by considering four steps; word cloud generator, acronym2word aggregation, false cognates removal, and domain knowledge stop-word removal. The result for the context-based text preprocessing had been forwarded into feature extractions with n-gram and TF-IDF approaches for the classifications. The result shows that our proposed framework has a higher Precision, Recall, F1-Score than Cor-TP. Furthermore, as explained in Tables 4, 5, and 6, the proposed framework shows more highlighted sections than Cor-TP, indicating better results. The inclusion of context-based text preprocessing showed performance improvement on majority classifiers in about 3% with the combinations of n-gram features.

There are several limitations to this study. First, the creation of the contextual-based text preprocessing is based on the word-cloud analysis results along with the experts' opinions. In the future, it is necessary to develop an approach to automatically add new words in the contextual text preprocessing. Second, the training data contain imbalanced class which may affect the performance accuracy. This study has no consideration on the imbalanced data and leave the issue for future work. Third, the class label involved the decision mainly on the experts meanwhile the complaint labels might be flexibly changed regarding the local needs. Hence, future work could consider unsupervised learning to group the relevant classes prior to the classification approach. At last, the feature extraction is limited to three (3) for the n-gram features while the local context might provide richer information with higher numbers. It is possible to extend this framework to other languages with the similar structure and semantics with Indonesian language (for example, Malay, or Tagalog). In addition, other variant of feature extraction would be interesting for future exploration.

REFERENCES

[1] S. Unger, M. Rollins, A. Tietz, and H. Dumais, "iNaturalist as an engaging tool for identifying organisms in outdoor activities," *J. Biol. Educ.*, vol. 55, no. 5, pp. 537–547, 2020, doi: 10.1080/00219266.2020.1739114.

[2] M. Levy and M. Germonprez, "The potential for citizen science in information systems research," *Commun. Assoc. Inf. Syst.*, vol. 40, no. 1, pp. 22–39, 2017, doi: 10.17705/1cais.04002.

[3] Y. Lin, "A comparison of selected western and Chinese smart governance: The application of ICT in governmental management, participation and collaboration," *Telecommun. Policy*, vol. 42, no. 10, pp. 800–809, Nov. 2018, doi: 10.1016/j.telpol.2018.07.003.

[4] M. Anshari and S. A. Lim, "E-government with big data enabled through smartphone for public services: Possibilities and challenges," *Int. J. Public Admin.*, vol. 40, no. 13, pp. 1143–1158, Nov. 2017, doi: 10.1080/01900692.2016.1242619.

[5] T. A. Oktariyanda and T. Rahaju, "E-government strategy of Surabaya city government through e-rt / rw to improve the quality of public service," *J. Phys., Conf. Ser.*, vol. 953, Jan. 2018, Art. no. 012161, doi: 10.1088/1742-6596/953/1/012161.

[6] A. R. Ziadi, B. Supriyono, and A. F. Wijaya, "The effectiveness of information system in public complaint service: An implementation of e-government based on Jakarta smart city applications," *Global J. Manage. Bus. Res., Admin. Manage.*, vol. 16, no. 8, pp. 53–57, 2016.

[7] R. Sarasati and E. D. Madyatmadja, "Evaluation of e-government LAKSA services to improve the interest of use of applications using technology acceptance model (TAM)," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 426, no. 1, Feb. 2020, Art. no. 012165, doi: 10.1088/1755-1315/426/1/012165.

[8] S. M. Zavattaro, P. E. French, and S. D. Mohanty, "A sentiment analysis of U.S. local government tweets: The connection between tone and citizen involvement," *Government Inf. Quart.*, vol. 32, no. 3, pp. 333–341, Jul. 2015, doi: 10.1016/j.giq.2015.03.003.

[9] Q. Mei and C. Zhai, "A mixture model for contextual text mining," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 649–655, doi: 10.1145/1150402.1150482.

[10] S. A. Macskassy, "Contextual linking behavior of bloggers: Leveraging text mining to enable topic-based analysis," *Soc. Netw. Anal. Mining*, vol. 1, no. 4, p. 355, 2011, doi: 10.1007/s13278-011-0026-8.

[11] J. Perkio, W. Buntine, and S. Perttu, "Exploring independent trends in a topic-based search engine," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Sep. 2004, pp. 664–668.

[12] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: An exploration of temporal text mining," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2005, pp. 198–207.

[13] D. Sebastian and K. A. Nugraha, "Text normalization for Indonesian abbreviated word using crowdsourcing method," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Jul. 2019, pp. 529–532, doi: 10.1109/ICOIACT46704.2019.8938463.

[14] E. Elakiya and N. Rajkumar, "Designing preprocessing framework (ERT) for text mining application," in *Proc. Int. Conf. IoT Appl. (ICIOT)*, May 2017, pp. 1–8, doi: 10.1109/ICIOTA.2017.8073613.

[15] T. Hofmann, "Probabilistic latent semantic analysis," 2013, *arXiv:1301.6705*.

[16] R. Andryani, E. S. Negara, and D. Triadi, "Social media analytics: Data utilization of social media for research," *J. Inf. Syst. Informat.*, vol. 1, no. 2, pp. 193–205, Sep. 2019, doi: 10.33557/journalisi.v1i2.23.

[17] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics—Challenges in topic discovery, data collection, and data preparation," *Int. J. Inf. Manage.*, vol. 39, pp. 156–168, Apr. 2018, doi: 10.1016/j.ijinfomgt.2017.12.002.

[18] T. Brandt, J. Bendler, and D. Neumann, "Social media analytics and value creation in urban smart tourism ecosystems," *Inf. Manage.*, vol. 54, no. 6, pp. 703–713, Sep. 2017, doi: 10.1016/j.im.2017.01.004.

[19] D. Gerrard, M. Sykora, and T. Jackson, "Social media analytics in museums: Extracting expressions of inspiration," *Museum Manage. Curatorship*, vol. 32, no. 3, pp. 232–250, Mar. 2017, doi: 10.1080/09647775.2017.1302815.

[20] E. A. Jensen, "Putting the methodological brakes on claims to measure national happiness through Twitter: Methodological limitations in social media analytics," *PLoS ONE*, vol. 12, no. 9, Sep. 2017, Art. no. e0180080, doi: 10.1371/journal.pone.0180080.

[21] Y.-C. Chang, C.-H. Ku, and C.-H. Chen, "Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor," *Int. J. Inf. Manage.*, vol. 48, pp. 263–279, Oct. 2019, doi: 10.1016/j.ijinfomgt.2017.11.001.

[22] W. He, X. Tian, Y. Chen, and D. Chong, "Actionable social media competitive analytics for understanding customer experiences," *J. Comput. Inf. Syst.*, vol. 56, no. 2, pp. 145–155, 2016, doi: 10.1080/08874417.2016.1117377.

[23] E. D. Madyatmadja, J. Olivia, and R. F. Sunaryo, "Priority analysis of community complaints through e-government based on social media," *Int. J. Recent Technol. Eng.*, vol. 8, no. 3, pp. 1–5, 2019, doi: 10.35940/ijrte.C5011.098319.

[24] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "A survey of text mining in social media: Facebook and Twitter perspectives," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 2, no. 1, pp. 127–133, 2017, doi: 10.25046/aj020115.

[25] M. Thangaraj and M. Sivakami, "Text classification techniques: A literature review," *Interdiscipl. J. Inf., Knowl., Manage.*, vol. 13, pp. 117–135, May 2018, doi: 10.28945/4066.

[26] A. V. Mantaring, M. A. P. Espinoza, and A. G. Gabriel, "Complaint management in the public sector organization in the Philippines," *Public Policy Admin. Res.*, vol. 9, no. 2, pp. 1–15, Feb. 2019, doi: 10.7176/ppar/9-2-03.

[27] M. M. Skoric, Q. Zhu, D. Goh, and N. Pang, "Social media and citizen engagement: A meta-analytic review," *New Media Soc.*, vol. 18, no. 9, pp. 1817–1839, Jul. 2016, doi: 10.1177/1461444815616221.

[28] A. Haro-de-Rosario, A. Sáez-Martín, and M. del Carmen Caba-Pérez, "Using social media to enhance citizen engagement with local government: Twitter or Facebook?" *New Media Soc.*, vol. 20, no. 1, pp. 29–49, Jan. 2018, doi: 10.1177/1461444816645652.

[29] R. Díaz-Díaz and D. Pérez-González, "Implementation of social media concepts for e-government: Case study of a social media tool for value co-creation and citizen participation," *J. Organizational End User Comput.*, vol. 28, no. 3, pp. 104–121, Jul. 2016, doi: 10.4018/JOEUC.2016070107.

[30] A. N. Kasiwi and A. Nurmandi, "The crowdsourcing model of social media in Surabaya's municipality, Indonesia," *Int. J. Manage., Innov. Entrepreneurial Res.*, vol. 4, no. 2, pp. 19–28, Jan. 2019, doi: 10.18510/ijmier.2018.423.

[31] H. K. Liu, "Crowdsourcing government: Lessons from multiple disciplines," *Public Admin. Rev.*, vol. 77, no. 5, pp. 656–667, Jul. 2017, doi: 10.1111/puar.12808.

[32] F. F. Mailoa, "Metode term frequencies untuk penelitian kesehatan di Twitter?: Studi pada tweet berbahasa Indonesia terkait obesitas," *Berita Kesehatan Masyarakat*, vol. 35, no. 4, pp. 1–4, 2019.

[33] A. F. Hadi, D. Bagus, and M. Hasan, "Text mining pada media sosial Twitter studi kasus?: Masa tenang pilkada DKI 2017 putaran 2," Seminar Nasional Matematika dan Aplikasinya, Universitas Airlangga, Surabaya, Indonesia, Oct. 2017.

[34] E. E. Pratama and R. L. Atmi, "A text mining implementation based on Twitter data to analyse information regarding corona virus in Indonesia," *J. Comput. Soc.*, vol. 1, no. 1, pp. 91–100, 2020.

[35] N. Yusliani, R. Primartha, and M. Diana, "Multiprocessing stemming: A case study of Indonesian stemming," *Int. J. Comput. Appl.*, vol. 182, no. 40, pp. 15–19, Feb. 2019, doi: 10.5120/ijca2019918476.

[36] Y. L. Arisanti, "Penggunaan akronim dan singkatan dalam media sosial Facebook di kalangan remaja SMA plus multazam," *Jurnal Literasi*, vol. 2, no. 2, pp. 104–112, 2018.

[37] M. Kaity and V. Balakrishnan, "An integrated semi-automated framework for domain-based polarity words extraction from an unannotated non-english corpus," *J. Supercomput.*, vol. 76, no. 12, pp. 9772–9799, Dec. 2020, doi: 10.1007/s11227-020-03222-0.

[38] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets responding to the Indonesian government's handling of COVID-19: Sentiment analysis using SVM with normalized poly kernel," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 2, p. 112, Oct. 2020, doi: 10.20473/jisebi.6.2.112-122.

[39] A. Librian, "High quality stemmer library for Indonesian Language (Bahasa)," GitHub, 2017.

[40] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[41] I. M. A. Agastya, "Pengaruh stemmer bahasa Indonesia terhadap peforma analisis sentimen terjemahan ulasan film," *Jurnal Tekno Kompak*, vol. 12, no. 1, p. 18, Feb. 2018, doi: 10.33365/jtk.v12i1.70.

[42] J. van Draanen, H. D. Tao, S. Gupta, and S. Liu, "Geographic differences in cannabis conversations on twitter: Infodemiology study," *JMIR Public Health Surveill.*, vol. 6, no. 4, pp. 1–10, 2020, doi: 10.2196/18540.

[43] L. Mesimova, "Colloquial words and expressions. Slang. Styles in written communication. Business communication," *Web Scholar*, vol. 4, no. 3, pp. 3–5, 2018.

[44] N. Roslidah and I. Komara, "Culture differences of Indonesia ethnic minorities in non-verbal communication," *Jurnal Studi Komunikasi*, vol. 1, no. 1, pp. 6–18, Mar. 2017, doi: 10.25139/jsk.v1i1.60.

[45] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *Proc. 47th Hawaii Int. Conf. Syst. Sci.*, Jan. 2014, pp. 1833–1842, doi: 10.1109/HICSS.2014.231.

[46] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale, "SparkClouds: Visualizing trends in tag clouds," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 6, pp. 1182–1189, Nov./Dec. 2010, doi: 10.1109/TVCG.2010.194.

[47] W. Ning, S. Lin, J. Zhou, Y. Guo, Y. Zhang, D. Peng, W. Deng, and Y. Xue, ‘‘WocEA: The visualization of functional enrichment results in word clouds,’’ *J. Genet. Genomics*, vol. 45, no. 7, pp. 415–417, Jul. 2018, doi: 10.1016/j.jgg.2018.02.008.

[48] R. L. Vislay-Wade and T. F. Scott, ‘‘Word cloud for hospitalization admission diagnoses in patients with multiple sclerosis,’’ *Multiple Sclerosis Rel. Disorders*, vol. 48, Feb. 2021, Art. no. 102681, doi: 10.1016/j.msard.2020.102681.

[49] I. N. Dewi and R. Nurcahyo, ‘‘Word cloud result of mobile payment user review in Indonesia,’’ in *Proc. IEEE 7th Int. Conf. Ind. Eng. Appl. (ICIEA)*, Apr. 2020, pp. 989–992, doi: 10.1109/ICIEA49774.2020.9102048.

[50] S. Deutsch, J. Schrammel, and M. Tscheligi, ‘‘Comparing different layouts of tag clouds: Findings on visual perception,’’ in *Human Aspects of Visualization*. Berlin, Germany, 2011, pp. 23–37.

[51] Y. Jin, ‘‘Development of word cloud generator software based on Python,’’ *Proc. Eng.*, vol. 174, pp. 788–792, Jan. 2017, doi: 10.1016/j.proeng.2017.01.223.

[52] F. O. Olùbódé-Sáwè, ‘‘Digital communication in indigenous languages,’’ in *Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction*. Hershey, PA, USA: IGI Global, 2010, pp. 564–578.

[53] F. Bond *et al.*, *Editors Program Committee Local Committee*, vol. 33, no. GWC. Indonesia: Linguistik Indonesia, 2018.

[54] S. Nakov, P. Nakov, and E. Paskaleva, ‘‘Cognate or false friend? Ask the web!’’ in *Proc. 1st Int. Workshop Acquisition Manage. Multilingual Lexicons*, Borovets, Bulgaria, 2007, pp. 55–62.

[55] D. S. Indraloka and B. Santosa, ‘‘Penerapan text mining untuk melakukan clustering data tweet shopee Indonesia,’’ *Jurnal Sains Seni ITS*, vol. 6, no. 2, pp. 6–11, Sep. 2017, doi: 10.12962/j23373520.v6i2.24419.

[56] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, ‘‘Improving text preprocessing for Student complaint document classification using sastrawi,’’ *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 874, no. 1, Jun. 2020, Art. no. 012017, doi: 10.1088/1757-899X/874/1/012017.

[57] B. Hou, F. Qi, Y. Zang, X. Zhang, Z. Liu, and M. Sun, ‘‘Try to substitute: An unsupervised Chinese word sense disambiguation method based on HowNet,’’ in *Proc. 28th Int. Conf. Comput. Linguistics*, 2021, pp. 1752–1757, doi: 10.18653/v1/2020.coling-main.155.

[58] R. Mahendra, H. Septiantri, H. A. Wibowo, R. Manurung, and M. Adriani, ‘‘Cross-lingual and supervised learning approach for Indonesian word sense disambiguation task,’’ in *Proc. 9th Global WordNet Conf.*, Jan. 2018, pp. 245–250.

[59] M. Schonlau and N. Guenther, ‘‘Text mining using n-grams,’’ *SSRN Electron. J.*, vol. 17, no. 4, pp. 866–881, 2016, doi: 10.2139/ssrn.2759033.

[60] M. Albathan, Y. Li, and A. Algarni, ‘‘Enhanced n-gram extraction using relevance feature discovery,’’ in *Proc. Australas. Joint Conf. Artif. Intell.*, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2013, pp. 453–465, doi: 10.1007/978-3-319-03680-9_46.

[61] A. Tripathy, A. Agrawal, and S. K. Rath, ‘‘Classification of sentiment reviews using n-gram machine learning approach,’’ *Expert Syst. Appl.*, vol. 57, pp. 117–126, Sep. 2016, doi: 10.1016/j.eswa.2016.03.028.

[62] N. Agarwal, G. Sikka, and L. K. Awasthi, ‘‘Enhancing web service clustering using length feature weight method for service description document vector space representation,’’ *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113682.

[63] N. Nofriani, ‘‘Comparations of supervised machine learning techniques in predicting the classification of the household’s welfare status,’’ *J. Pekommas*, vol. 4, no. 1, p. 43, Apr. 2019, doi: 10.30818/jpkm.2019.2040105.

[64] M. Mohanapriya and J. Lekha, ‘‘Comparative study between decision tree and KNN of data mining classification technique,’’ *J. Phys., Conf. Ser.*, vol. 1142, p. 12011, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012011.

[65] M. Li, P. Xiao, and J. Zhang, ‘‘Text classification based on ensemble extreme learning machine,’’ 2018, *arXiv:1805.06525*.

[66] N. Bahrawi, ‘‘Sentiment analysis using random forest algorithm-online social media based,’’ *J. Inf. Technol. Utilization*, vol. 2, no. 2, pp. 29–33, 2019.

[67] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. R. Sinha, ‘‘An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach,’’ *IEEE Access*, vol. 9, pp. 67488–67500, 2021, doi: 10.1109/ACCESS.2021.3075140.

[68] D. Singh and S. Singh, ‘‘Realising transfer learning through convolutional neural network and support vector machine for mental task classification,’’ *Electron. Lett.*, vol. 56, no. 25, pp. 1375–1378, Dec. 2020, doi: 10.1049/el.2020.2632.

[69] I. S. Al-Mejibli, J. K. Alwan, and D. H. Abd, ‘‘The effect of gamma value on support vector machine performance with different kernels,’’ *Int. J. Electr. Comput. Eng.*, vol. 10, no. 5, p. 5497, Oct. 2020, doi: 10.11591/ijece.v10i5.pp5497-5506.

[70] M. Z. H. Jesmeen, J. Hossen, S. Sayeed, C. K. Ho, K. Tawsif, A. Rahman, and E. Arif, ‘‘A survey on cleaning dirty data using machine learning paradigm for big data analytics,’’ *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 10, no. 3, pp. 1234–1243, 2018, doi: 10.11591/ijeecs.v10.i3.pp1234-1243.

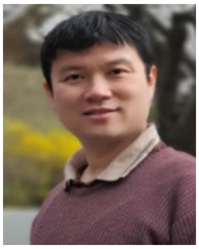
[71] A. Maulana and A. Romadhony, ‘‘Domain adaptation for part-of-speech tagging of Indonesian text using affix information,’’ *Proc. Comput. Sci.*, vol. 179, no. 2020, pp. 640–647, 2021, doi: 10.1016/j.procs.2021.01.050.

[72] D. W. Castro, E. Souza, D. Vitório, D. Santos, and A. L. I. Oliveira, ‘‘Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties,’’ *Appl. Soft Comput.*, vol. 61, pp. 1160–1172, Dec. 2017, doi: 10.1016/j.asoc.2017.05.065.

[73] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, ‘‘Credit card fraud detection using AdaBoost and majority voting,’’ *IEEE Access*, vol. 6, pp. 14277–14284, 2018, doi: 10.1109/ACCESS.2018.2806420.

[74] R. Atallah and A. Al-Mousa, ‘‘Heart disease detection using machine learning majority voting ensemble method,’’ in *Proc. 2nd Int. Conf. New Trends Comput. Sci. (ICTCS)*, Oct. 2019, pp. 1–6, doi: 10.1109/ICTCS.2019.8923053.

EVARISTUS DIDIK MADYATMADJA (Member, IEEE) received the master’s degree in computer science from Gadjah Mada University, Yogyakarta, Indonesia, in 2005, and the Ph.D. degree in computer science from Bina Nusantara University, Jakarta, Indonesia, in 2019. He is currently an Associate Professor with the School of Information Systems, Bina Nusantara University. He is the author of more than 60 articles. He is an Editor-in-Chief of the journal of *Computer, Mathematics and Engineering Applications* (Comtech). In the last few years, he has conducted smart city research in several cities in Indonesia. His research interests include database, UI/UX, big data, data analytics, and business intelligence.



BERNARDO NUGROHO YAHYA received the B.S. degree in industrial engineering from Petra Christian University, the M.S. degree in information system engineering from Dongseo University, and the Ph.D. degree in industrial engineering from Pusan National University. He is currently a Full Professor with the Industrial and Management Engineering Department, Hankuk University of Foreign Studies, South Korea. He has been working on various industry business consulting and engineering projects with Korean companies. His current research includes statistical pattern recognition, machine learning, business process intelligence, and data analytics

CRISTOFER WIJAYA received the bachelor’s degree in information systems from Bina Nusantara University, Jakarta, Indonesia.

From 2020 to 2021, he was a Junior Researcher with the School of Information System, Bina Nusantara University. His research interests include data science, machine learning, and information systems.