Revue d'Intelligence Artificielle
Vol. 38, No. 1, February, 2024, pp. 63-72

Journal homepage: http://iieta.org/journals/ria

Textual Analysis for Public Sentiment Toward National Police Using CRISP-DM Framework

Latifa Z.S. Sudar*, Joash L. Imbenay, Indra Budi, Amanah Ramadiah, Prabu K. Putra, Aris B. Santoso

Faculty of Computer Science, Universitas Indonesia (UI), Jakarta 10430, Indonesia

Corresponding Author Email: latifa.zahra@ui.ac.id

https://doi.org/10.18280/ria.380107

Received: 15 September 2023
Revised: 5 December 2023
Accepted: 12 December 2023
Available online: 29 February 2024

Keywords:
Lexicon, machine learning, CRISP-DM, POLRI, sentiment analysis, topic modeling

ABSTRACT

Nowadays, public opinion toward the National Police's (POLRI) image is deteriorating. With the explosive growth of social media in Indonesia, opinions on POLRI-related present-day issues on Twitter easily go viral, influencing sentiments among individuals regarding Indonesian law enforcement. Negative sentiments, at some point, may lead to the undervaluation of law enforcement and the failure of the legal system. Therefore, sentiment analysis on Twitter is essential for gaining considerable insights into public views and attitudes on POLRI-related topics. This research is to determine the most effective approaches between Lexicon, a natural language processing method that relies on a corpus, and machine learning, which contains Naive-Bayes, Support Vector Machine (SVM), Random Forest, and Logistic Regression (LR). These approaches have differences in classification types: probability and linearity. To organize the research process, the Cross-Industry Standard Process for Data Mining (CRISP-DM) Framework, which comprises five data mining activities, was employed. The confusion matrix was used as the model performance measurement, with Naive-Bayes emerging as the best among all the tested models. Additionally, the subjects related to POLRI were developed using topic modeling, generating three topics: street police or police station, police acknowledgment in neighborhood activities, and the activity of contacting the police.

1. INTRODUCTION

Since the early 2000s, the rapid rise of social media has resulted in substantial changes in human communication and connection. This is caused by the role of social media platforms in enabling users to broadcast interpersonal conversation to huge audiences, which expands social networking to a wider range of family and friends [1]. In Indonesia, as one of the social media platforms, Twitter has emerged as a prominent social media platform. Indonesia is placed sixth with 14.75 million Twitter users out of a total of 372.9 million users in 2023, according to We Are Social research.

Due to the growing number of social media users in Indonesia on Twitter, any type of tweet has the opportunity to go viral and become a national hot topic. Further, Twitter has been extensively documented as a hotspot for the spreading of narratives [2]. For example, three significant shifts from online narratives to offline activity occurred in 2019: the post-election riots, the student uprising against the alleged racist treatment of Papua students in Surabaya, and the harm done by Jakarta students to the antigraft law.

The Indonesian law enforcement system is one of the subjects that Indonesian Twitter users frequently draw attention to, among other things. The objections and grievances regarding law enforcement activities have the potential to affect public opinion. When the public opinion is negative, it can incite public outrage and increase public suspicion of the law enforcement [3]. Moreover, the public's belief in unethical law enforcement might erode respect for Indonesian systems of law. Undervalued law systems result in a decline in collaboration, an increase in crime, and a failure of the justice system.

Law enforcement is known to be related to government organizations such as the National Police (POLRI). According to Constitution No. 2 for the year 2022, POLRI is national authorities given roles in maintaining public discipline, protection, and serving people to safeguard domestic security. POLRI are assigned as a shelter to reduce internal security dilemmas by building relationships between the police and the Indonesian civilian community, as well as any related parties from other countries that have legal agreements with Indonesia.

Before the formation of the POLRI, the function of maintaining national security was occupied by a military force called the ABRI. Then, in 1988, POLRI became separated from ABRI and given its own role to serve in Indonesia as civilian police [4]. While still considered to be part of ABRI, POLRI is strongly associated with a violent and militaristic approach. Despite already being independent from ABRI, the image of violence is still attached. This was worsened by several acts of indiscipline by officers on duty [5]. Evidence indicates that the public's perception of the role and function of POLRI in maintaining order and security is not yet in line with public expectations [6]. Hence, the opinion of the public regarding the POLRI image are deteriorating on a daily basis.

As POLRI relies on the trust and cooperation of the Indonesian people to fulfil its mission, the effectiveness and credibility of POLRI are closely tied to public sentiment. Thus, in today's digital age of internet communication, assessing public sentiment toward POLRI has become critical, as positive or negative public opinion can have a tremendous impact on the performance and reputation of POLRI. Therefore, research is required to analyze public sentiment toward POLRI.

As a subfield of natural language processing (NLP) and machine learning, sentiment analysis provides a reliable methodology for analyzing and measuring public sentiment expressed on social media such as Twitter [7]. Data from Twitter was regarded as a convenience sample because it provides non-constructed samples of non-random human behaviors as the population [8]. Due to the fact that tweets often reveal user sentiment on contentious topics, Twitter has recently been the target of extensive research [9]. Researchers can acquire significant insights into public views, attitudes, and opinions concerning diverse topics, including government institutions such as POLRI, through the application of sentiment analysis techniques to massive datasets of Twitter data. This information can be advantageous for identifying areas in which POLRI is seen favorably and areas in which they may need to address issues, enhance methods for interactions, or take corrective action. Thus, this research aims to study the public sentiment towards POLRI by observing the tendency of Indonesian citizens through Twitter using several techniques to find the most suitable algorithm. The result of the research is expected to accomplish the following:

(1) Identify The Whole Sentiments: To assess the overall sentiment of POLRI-related tweets, which can later be categorized as positive and negative.

(2) Find Key Topics: To narrow down and analyze key topics or issues that are typically mentioned in POLRI-related tweets on Twitter, revealing the public's key tensions and interests.

Furthermore, while sentiment analysis acknowledges public opinion regarding the performance of POLRI through its semantic meaning from Twitter as microblog medium, this brevity presents challenges and opportunities. Results from this research will provide a broader comprehension of public perceptions of POLRI as represented on Twitter. This study contributes to an improved understanding of how relationships develop between law enforcement and the communities they serve. Additionally, it has the potential to assist POLRI in improving its communication skills, accommodating public concerns, and ultimately establishing a more positive and empowered connection with the Indonesian public.

2. PROPOSED METHOD

This research employs the cycle method of Cross Industry Standard Process for Data Mining (CRISP-DM) as a framework. CRISP-DM is a systematic method for dependent data projects that supports collaboration and assures data-driven solutions satisfy business objectives. CRISP-DM is commonly used for handling and executing data mining and machine learning projects, which consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, as shown in Figure 1 [10, 11].

The implementation of CRISP-DM is segmented in accordance with the task definition [10]. The Business Understanding phase determines project objectives and requirements while simultaneously defining an accurate idea of the problem domain. For further insights into the dataset, the Data Understanding phase contains data collection, exploration, and preliminary assessment. The purpose of Data Preparation phase is to clean, process, identify the relevant data for analysis, and label the data according to its sentiment. The Modeling phase comprises the selection and execution of relevant algorithmic methods for machine learning to create prediction models, and extract topics to identify current issues. The model's performance is then reviewed at the Evaluation phase to ensure that it meets the business objectives. Finally, the Deployment phase implements the model by integrating it into the operational environment.

The proposed method can be prescribed based on the structure described above as shown in Figure 2. The activities at data understanding phase are correlated to studies linked to the case and the discovery of tools to be employed for each phase. Following that, data are collected and interpreted, including data recollection if the initial data is insufficient for analysis. Following data collection, the subsequent activity is data preprocessing, which is useful for cleaning the data obtained, and data annotation. It is used to label the data for further analysis. The modeling phase is then completed by the use of several machine learning algorithms for further sentiment analysis and topic modeling in addition to topic analysis. Evaluation phase is held by performing confusion matrix and using its measurement to determine the proper model for this study. Finally, as part of deployment phase of this study, an analysis is conducted to draw conclusions on the analysis conducted prior to correspondence with the case study.

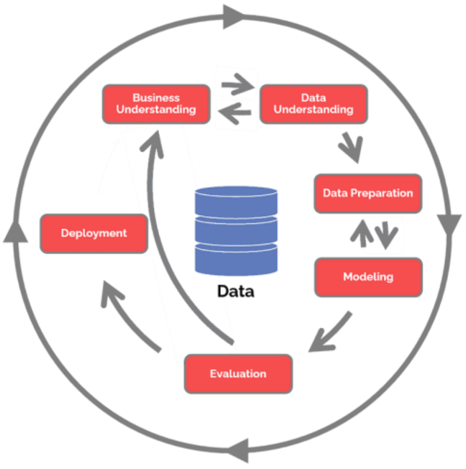
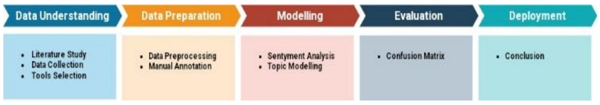


Figure 1. CRISP-DM cycle



Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
• Literature Study • Data Collection • Tools Selection	• Data Preprocessing • Manual Annotation	• Sentyment Analysis • Topic Modelling	• Confusion Matrix	• Conclusion

Figure 2. CRISP-DM activities for text mining

2.1 Information extraction

Information extraction in text mining is considered as a key process with the aim of finding insightful information by identifying specific pieces of information from text, such as entities or relationships between entities from unstructured or semi-structured texts [7]. Entity recognition techniques are employed to extract specifically named entities such as person names, organization names, or locations from the text, which can be valuable for the analysis. The initial step requires transforming the unstructured, amorphous text files into a structured database, which allows the implementation of data mining techniques to extract valuable information and interesting patterns. In a relational database, data are organized in tables with rows and columns to retrieve the required information, since structured queries requires the column and table names to be identified. However, the adversity of information extraction from unstructured data lies in the process since the specific patterns in the text are difficult to identify.

2.2 Data preprocessing

Data preprocessing in text analysis is useful for enhancing the accuracy and effectiveness of the analysis [12, 13]. This step includes cleaning the text by removing irrelevant characters, symbols, and numbers, as well as handling issues like capitalization and spelling variations. The procedure entails deleting stopword, which is useful for removing words with less significant meanings in the context of the study. Furthermore, techniques like stemming or lemmatization are applied to reduce words to their root form, which manages morphological variations, ensuring consistency and improving the accuracy of analysis. The goal of this process is to obtain a cleaner and more structured dataset, which enables more accurate and meaningful insights for analysis such as sentiment analysis or topic modeling.

2.3 Sentiment analysis

Sentiment analysis is the study of public opinion, feelings, sentiments, and attitudes about various objects using written language [14]. The major assignment in sentiment analysis is to classify whether a given opinion extracted from the public space such as a miniblog is positive or negative, as well as the depth of the sentiment. In addition to that, sentiment analysis is vital for data analysis to assist the party concerned in making decisions. The information used as input for sentiment analysis is often textual in nature which includes reviews, social media posts about news articles, and comments on internet forums. Sentiment analysis algorithms process this textual data and generate relevant insights through choices of methodologies, for example, machine learning and deep learning. A classification resulted from a sentiment analysis process reveals which of the sentiments stated in the text input is positive, negative, or neutral. Particularly, sentiment analysis methods additionally include a numerical sentiment score or confidence level, which provides an exact indicator for sentiment measurement.

In the sentiment analysis field, classification methods have evolved many strategies for classifying data. The most common classification method is when machine is used as a tool to train certain dataset. Another classification method is when Lexicon is used as a base to show the sentiment classes, positive and negative.

A machine-based approach, well known as machine learning, needs to be trained with large amounts of data. As a branch of Artificial Intelligence, machine learning focuses on algorithms and techniques development that enable machines to learn from data samples with the aim of producing accurate predictions from unknown data. Machine Learning is used for analyzing given data and identifying patterns or rules within it. This allows computers to make predictions or take actions based on new, unseen data [8, 15]. Among the existing machine learning algorithms, there are several popular algorithms that are often used for text analysis and sentiment analysis research. The algorithm used for classification is considered to be supervised. These supervised machine learning algorithms are divided into two types of approaches: probabilistic classifiers and linear classifiers [16] as shown at Figure 3.

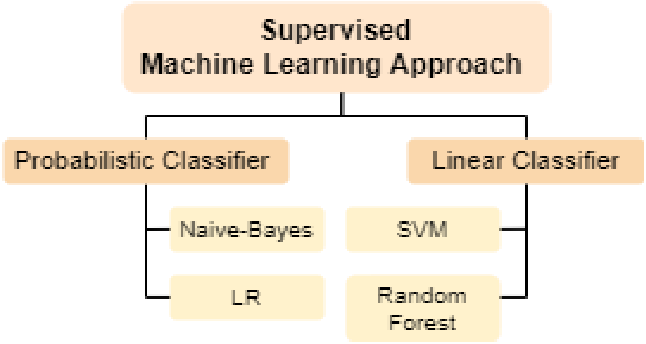


Figure 3. Machine learning classification

To classify data, the probability classifier employs mixture models. It treats each class as a component of the mixture. One of the models categorized as probabilistic classifiers is Naive-Bayes. Naive-Bayes is a probabilistic machine learning algorithm for classification based on Bayes' theorem. It assumes independence between the features by calculating the conditional probability of a given sample belonging to each class based on the observed features, and then predicts the class with the highest probability [16, 17]. Furthermore, Logistic Regression (LR) method is also categorized as a probabilistic classifier. LR method employs probability of an instance being classified into a specific class using the sigmoid function by estimating coefficients from the sample data that maximize the similarity to predict the probabilities of the unknown data [16, 18]. While Naive-Bayes is a generative model, LR is described as a discriminative model, and its results are both interpretable and efficient, thereby making it suitable for binary classification issues.

The linear classification algorithm is designed for separating items with similar feature values into units. The classification preference was made by a linear classifier based on the value of a linear set of features. The example of popular linear classifier are Support Vector Machine (SVM) and Random Forest. SVM is a machine learning algorithm used for classification and regression. It works by constructing a model that can separate two classes using a hyperplane with the largest margin between the classes. The hyperplane is selected in such a way that it maximizes the distance from the nearest data points to the separating line [16, 19]. Random Forest is an ensemble learning method known as the extension of decision tree which combines decision trees by selecting random subsets of the training data and features, then aggregates the results from each tree using a voting mechanism for classification or averaging for regression [16].

Meanwhile, Lexicon Based approach is often used to obtain the sentiment classification of a text, which is categorized as unsupervised classifier [20, 21]. Lexicon Based approach converts each sentence directly into a score without data training. These classifiers work by identifying each word of the text and converting each word into numeric form based on the Lexicon attribute which contains weight of words, to discover the whole text sentiment. Lexicon Based approach does not require any data sample since the conversion stage utilizes the weighted word that had been developed as a dictionary. The main process of Lexicon Based approach involves simple mathematical equation that reveals the sum of the full text as positive identified as positive opinions, and negative result as negative opinions.

To determine the best performance of the various methods that have been described, this research proposes an examination of machine learning approaches and Lexicon based approach. To be precise, a comparison between Naive-Bayes, LR, SVM, Random Forest, and Lexicon was held to discover the most suitable algorithm regarding this subject.

2.4 Topic modeling

A type of statistical approach designated for topic modeling has been used to determine the root ideas or topics in a population of natural texts [22]. It has been frequently employed in performing content distribution or identifying the key concepts in an enormous quantity of text. One of the models widely used as a topical generative probabilistic model is Latent Dirichlet Allocation (LDA).

LDA model predicates the idea that each document consists of a variety of themes, whereas all the topics are combinations of words [23]. Discovering the topic distributions and keyword patterns that best depict the themes while clarifying the words in the documents are the two primary targets of LDA. This method for topic modeling consists of two stages. The initialization phase aims to establish the k-th extractable topics for initializing the topic key and text topic distributions at random. Afterwards, the iterative process phase is held, which consists of two main tasks. The first task is to calculate the probability of a word belonging to a subject for every single word within the whole document. Based on these results, words are then reassigned to topics, and topic distributions are updated. The results, which consist of words, are generated after the algorithm converges. Thus, the word and topic distributions remain unchanged after numerous iterations. A high percentage of words inside a topic reveal the theme of the topic. The variety of topics in documents corresponds to the uniqueness of the topics [24]. The probability of the document formulated as:

$$p(D \mid \alpha,\beta) = \prod_{d=1}^{M} \int p\,(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p\,(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn},\beta) \right) d\theta_d$$

where, $D = \{d_1, d_2, \dots, d_m\}$ is defined as the dataset of documents that contain M documents, and each document d is composed of N words defined as $w = \{w_{d1}, w_{d2}, \dots, w_{dn}\}$. α defined as parameters of Dirichlet prior topic, and β defined as the distribution of words across topics, generated based on the Dirichlet distribution. The multinomial distribution for document d is written as θ_d, and the distribution of document level θ_d generates topic per document z_{dn} . The implementation of this formula shown at Figure 4, where it can be seen that the input from the system is dirichlet prior topic per document and word across topic. Further, the system output is words with a latency degree to the tune of the k value initiated earlier.

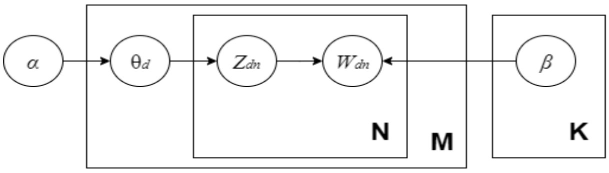


Figure 4. LDA model

2.5 Confusion matrix measurement

Confusion matrix measurement is often used in machine learning as an evaluation tool of the classification model performance, which provides a tabular representation of the predicted and actual class labels of a dataset. The matrix is constructed with four elements such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN) [25-27]. On positive prediction, the model generates either true positive (TP) or false positive (FP) labels. Negative prediction, on the other hand, generates both true negative (TN) and false negative (TN) labels. TP gives insight into the positive labels predicted correctly by the model as the positive classes, while FP represents the cases of the negative labels predicted as positive classes by the model. Moreover, TN reveals the negative labels predicted correctly as the negative classes, and FN exposes the cases of the positive labels predicted as negative classes by the model.

Confusion matrix provides calculation of various evaluation metrics such as accuracy, precision, recall, and F-score [26]. Accuracy means the measurement of all the correct prediction of the model. Precision calculates the amount of correct prediction with positive labels compared to all the predictions with positive classes. Recall calculates the amount of correct prediction with positive labels compared to all the actual positive labels. F-score is the harmonic means of precision and recall, supplying a balanced measure between the two.

2.6 Related study

There are several types of sentiment classification techniques, such as machine learning, Hybrid Based techniques, and Lexicon Based techniques, though this study only discusses machine learning and Lexicon Based techniques [28]. A Study from Ilmania et al. [29] focuses on optimizing a Lexicon Based techniques because the real data to test have no data sample as the requirement for performing supervised learning. To optimize the accuracy of sentiment classification, a combination of Lexicon and neural network is used in the study of sentiment analysis or text mining.

There is a study that uses neural network-based techniques on the sentiment Lexicon combining the multi-layer perceptron (MLP) and BiLSTM [30]. This method is claimed to be highly accurate, which results in up to 80% accuracy measured by F-score. Further, Pan et al. [31] shows a deep neural network by experimenting with four different structures with a result of 85.29% on the F-score from Connected Neural Network (CNN).

Research on sentiment analysis with machine learning has been conducted often. Previous research by Brazilian researchers resulted in Naïve-Bayes and SVM methods showing high accuracy of 0.82 and 0.84 with 0.819 and 0.821 of F-scores, respectively [32]. The SVM method shows a result of 0.873 from the F-score and 0.80 from the accuracy for sentiments detection. On the other hand, Naïve-Bayes method presented a result of 0.791 from the F-score and 0.727 from the accuracy. The application of the SVM method increases accuracy up to 8%, as it is enhancing the model. Research from Genç and Surer (2023) revealed that classification of click bait strategy on social media gains 85% of accuracy and F-score from LR, and 86% of accuracy and F-score from Random Forest, and the ensemble models gain accuracy and F-score to 93% [33].

Subsequently, topic modeling with LDA algorithm is used for classifying significant topics of texts. In the study by Huang et al. [22], the LDA algorithm generated eight different topics for crowdfunding reviews from the participants. This method has also been combined with Non-negative Matrix Factorization (NMF) to determine topics from Urdu text with the result of merging texts into a longer phase for better LDA [34].

3. RESEARCH METHOD

The elaboration of the CRISP-DM framework was used as the method for this research. Since the CRISP-DM framework's corresponding activities have been specified, the implementation of CRISP-DM in accordance with this research is shown in Figure 5.

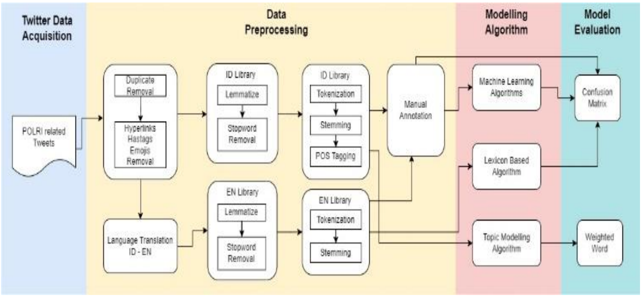


Figure 5. Research flow based on CRISP-DM framework

3.1 Data collection

Data are collected from data sources gathered through information extraction techniques. Data collection used in the analysis was obtained from Twitter application programming interface (API), which performs data crawling with certain keywords. In this analysis, we use "POLRI" or "Polisi" as the keyword of the search. To specify the search, we filtered only tweets that were sent in response to a news or thread regarding the Indonesian police or justice. The time frame of the data sample was between April 13th of 2023 and April 20th of 2023, with a total of 1159 tweets crawled. However, due to the irrelevant content of the tweets, the data were reduced to 800 tweets.

3.2 Data pre-processing

In text mining, data preprocessing refers to the fundamental procedure of cleaning, prepping, and conversing raw textual data into a structure designed for analysis. Data pre-processing is the step in the sentiment analysis to ensure that the data are in the right format and quality. Effective data preprocessing is critical because it enhances the quality of the input data and the effectiveness of the text mining steps that follow, including text classification, sentiment analysis, and topic modeling. Data preprocessing consists of several steps such as cleansing the data, transforming text to lowercase, tokenization which separates texts into single words or tokens, stemming or lemmatization which decreases words to their base form, and removing stop words. The step of cleaning the data is done by removing slang words and abbreviations, as well as removing punctuation. Moreover, addressing any inconsistencies or errors in the dataset is also crucial; this may include dealing with missing data, encoding difficulties, and addressing any distinctive noise or abnormalities of the original text. To accomplish this, a word replacement technique was employed, utilizing a slang word and abbreviation library developed by the author. The task was to identify and replace slang words and abbreviations with their corresponding standard counterparts or more formal equivalents.

After, tokenization is performed to separate the text into words in order to determine tokens to be used as features. This step helps in preparing the data for further analysis and feature extraction. Additionally, stopword removal was carried out to eliminate common words with insignificant meaning or contribute to the sentiment analysis as the words are removed from the sentences by code during this step to focus on more meaningful content. Subsequently, lemmatization is applied to reduce inflected words to their base or dictionary form, known as lemmas. This helps consolidate different forms of a word into a single representation, aiding in the accurate interpretation of sentiment. The cleaning, tokenization, stopword removal, and lemmatization processes were all implemented using Python, leveraging relevant libraries and techniques to ensure that the data were appropriately prepared for sentiment analysis. Python has several libraries for processing natural language. For this research, we employ the NLP library which contains NLTK, to perform tokenization, lemmatization, and stopword removal, and other libraries to perform data processing such as Pandas, Random, CSV, and Re.

After processing text into a cleaner dataset form, the distribution of the dataset is held. Thus, dataset separated into training dataset and testing dataset for machine learning method. While training set was used as exercise for machine to generate the classification model, the testing set was used as evaluation towards its performance on unknown data. A total of 70% of 800 data separated into training data and the rest for testing data for performing several machine learning algorithms. However, for Lexicon Based algorithms, all collected dataset is used for algorithm performance since we do not need to separate the datasets into training and testing.

3.3 Data annotation

The collected dataset that consists of natural language is labeled according to its semantic meaning in the sentiment classification, which is a positive and negative label. Though we found several tweets with neutral sentiment, we excluded the neutral label from this research, as we consider the neutral sentiment toward POLRI to mean the person behind the examined text has no tendency toward negative feelings. In this experiment, the collected dataset is then manually labeled for subsequent accuracy measurement. Thus, we classify the text as positive, as presented in Table 1.

Overall data collected and labeled in this study provides imbalanced dataset as captured in Figure 6, which respectively represent negative class and positive class.

Table 1. Example of two stage data annotations

Text		Sentiment	
ID	EN	Stage I	Stage II
@kompascom Bravo polisi bisa menemukan tersangka penampakan dalam video. KPK perlu belajar menemukan HM.	@kompascom Bravo police found culprit of appearance in the video. KPK should learn to find HM.	Positive	Positive
@detikcom Setuju pejabat Publik anti kritik ditegur netizen. Kritik harta bernilai bukan malapetaka harus lapor ke Polisi.	@detikcom Agree the public's functionary attitude is anti-criticism. Criticism is valuable and not something to be reported to the police.	Neutral	Positive
@CNNIndonesia Begitulah Polisi bermental budak, bukan bela benar malah bela jahat.	@CNNIndonesia Exactly slave mental police, defending evil instead of good	Negative	Negative

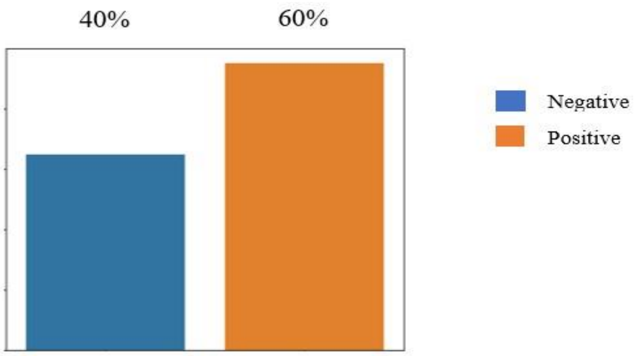


Figure 6. Imbalanced dataset

3.4 Sentiment analysis

In this research, sentiment analysis is performed using Python as a programming language. There are two main libraries that have been used to develop the algorithms, namely SKLearn and NLTK classification library. We used the algorithms to perform classification and topic modeling.

3.4.1 Classification algorithm

The appropriate algorithm was chosen to select a model that can effectively capture the underlying patterns in the data to make accurate classifications for sentiment prediction. In this research, Lexicon Based approach, and machine learning methods such as Naive-Bayes, LR, SVM, and Random Forest were selected. The outcome of the sentiment analysis in its entirety depends on the algorithm selection, which has a significant effect on the accuracy of sentiment forecasts. The first algorithm chosen for this study is Lexicon Based. The general steps involved in a Lexicon Based approach are text preprocessing, text translation, stopword removing, and word score as shown in Figure 7.

Text preprocessing was conducted as part of data preparation mentioned earlier. However, the step involving removing the stop word was excluded to ensure that the text was ready to be processed further because removing words can be damaging to the meaning of the text. Due to the lack of an Indonesian corpus for the topics, an official English corpus was selected for this analysis to enhance the weight accuracy of the words in the text. The English corpus used for this research is from the Natural Language Toolkit (NLT) library in Python. This library provides several corpus, namely Sentiwordnet and Wordnet. SentiWordNet contains a lexical resource for opinion mining with output values of positive, negative, and neutral. The use of wordnet helps machines understand synonyms, antonyms, and other relationships between words. Thus, text translation from Indonesia to English is needed.

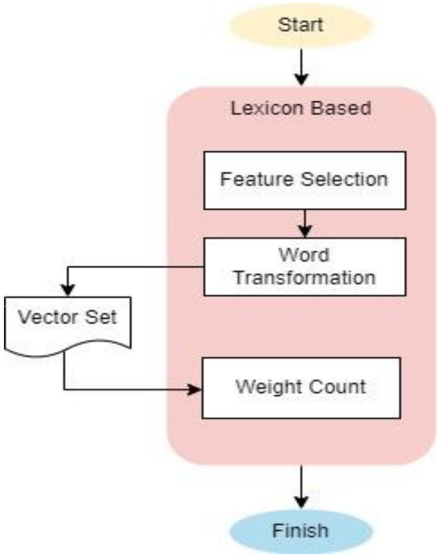


Figure 7. Steps in Lexicon based approach

After translating Indonesian text into English, stopword removal is performed to reduce the number of words to be scored based on the weight. The stopwords corpus was also provided by the NLT of the Python library. The last step is to assign sentiment scores to each word in the text using the selected Lexicon. Each word in the Lexicon is looked up and its associated sentiment score is retrieved. The score that was included from SentiWordNet can indicate whether the word is positive, negative, or neutral. However, to simplify the case, neutral classification on the dataset is considered similar to positive classification because the initial separation algorithm only observed negative classification and anything else was positive.

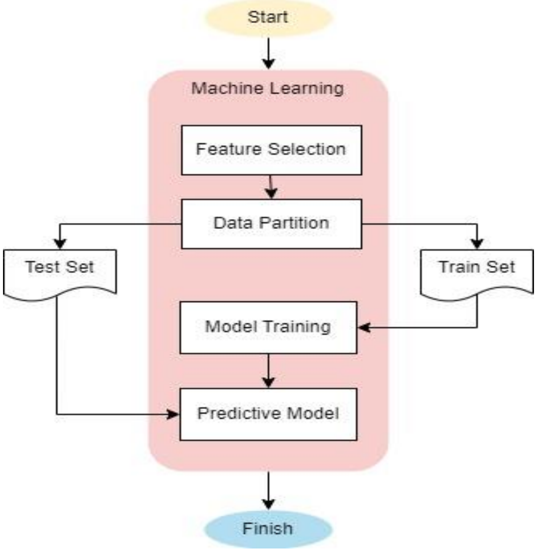


Figure 8. Steps in machine learning approach

The rest of the algorithms used for this study are obtained from supervised machine learning method. Machine learning algorithms consist of the major steps; they are text preparation, stopword removal, features selection, and training dataset. The first step was to prepare the text by cleaning, preprocessing, and splitting dataset as the part of data preparation mentioned earlier. The next step is to clean the data from insignificant words that were accounted for as stoppers because several words created bias for the analysis. Afterward, the relevant features are selected to be used as inputs to the classifier algorithm. This step has a significant impact on the classification task as the classifier assumes features. The last step was to evaluate the model after training the classifier with the training dataset, which independently take each feature into account to classify the class label, as shown in Figure 8.

3.4.2 Topic modeling

Subsequently, a topic modeling algorithm was performed, which also uses a library directly from Phyton, namely LDA. To perform this algorithm, the same text that has been crawled for sentiment analysis is used. The first step is to cleanse the text from abbreviations and slang words, as well as clean words without significant meaning. After that, the data are transferred into vector form to represent words as matrices using the vectorizer from Phyton. For this research, we bound the topics and words up to 10 to limit the running of the algorithm.

4. RESULT AND DISCUSSION

After executing the selected techniques, confusion matrix can be generated to evaluate the performance of each technique as summarized in Table 2.

The higher percentage of true positives (compared to the others) indicates that predictions of positive sentiment models are more reliable in identifying patterns of positive sentiment in the data. To achieve a more comprehensive view of the efficiency of the model across different points of view, it is crucial to include consideration of additional measures like precision, recall, and F1-score.

Table 2. Confusion matrix result

Technique	TP	FP	TN	FN
Lexicon	334	210	114	142
Naive-Bayes	137	32	57	14
SVM	113	41	48	38
Random Forest	134	67	27	12
LR	116	52	43	29

The result shows that Lexicon Based approach and machine learning method perform with different scores of precisions, recall, F-score, and accuracy. The results obtained differ from previous research, which stated that different techniques yield scores that are not significantly different. In contrast, the conducted experiment provides results of the best performance achieved by Naive-Bayes, followed by Random Forest, SVM, LR, and Lexicon Based approach when compared by F-score value. In the context of sentiment analysis using Twitter data, superior performance of a model is determined by achieving high precision and high recall which can signify the model's effectiveness in detecting the number of tweets with positive sentiment. It is crucial, however, to tailor the evaluation criteria so that they align with the specific objectives and requirements of the project or research at hand. As for this experiment, an F-score of 65.49% presented in Table 3 shows that the Lexicon model has enough equilibration on combination of precision and recall.

Table 3. Evaluation using Lexicon based and machine learning

Technique	Precision	Recall	Accuracy	F-Score
Lexicon	61.39	70.16	56.00	65.49
Naive-Bayes	81.06	90.72	80.83	8.62
SVM	73.37	74.83	67.08	74.09
Random Forest	66.67	91.78	67.08	77.23
LR	69.00	80.82	66.25	74.41

Meanwhile, the table below also shows the F-score of machine learning methods which are 85.62%, 74.09%, 77.23%, 74.41% for Naive-Bayes, SVM, Random Forest, and LR respectively, suggesting that the optimum equilibrium is performed by Naive-Bayes method. Thus, high value of F-score measurement ensures that the performance of the model works well within positive and negative classes, which suggests that the model can be used in this case study.

Beside F-score, a high accuracy value indicates model accuracy of predictions, while a low value suggests that the model is struggling to classify instances correctly. This experiment produces accuracy of 56% executed by Lexicon Based approach, 80.83% by Naive-Bayes, 67.08% by SVM and Random Forest, and 66.25% by LR. It means the ratio of the total number of correct predictions to the total number of predictions made by machine learning method were more improved compared to those generated by Lexicon Based approach. The high value of accuracy represents the performance of the model in correctly predicting all classes, which reveals that the model is considered effective in predicting the classes in this case study.

In this study, although Naive-Bayes algorithm yielded the most effective evaluation result compared with all the algorithms tested in this research, it does not suggest that the performance of other algorithms is lacking considering the amount of dataset to be trained and tested were low. The evaluation of SVM and Random Forest gives the exact same accuracy with slightly different F-score values, which indicates the performance of two classifiers show equal consistency in performance. The result of the evaluation can be improved by implementing cross validation or by hyperparameter tuning. However, for this research, it needs to be emphasized that all algorithms tested can serve as tools to identify sentiment within opinions on Twitter.

Nonetheless, it is important to note that F-score and accuracy may not provide a complete picture of model performance as the dataset used for the experiment is quite imbalanced due to the irrelevant amount of data crawled at the beginning of experiment [35]. Also, in the Lexicon Based approach, an official sentiment analysis corpus on Phyton called SentiWordNet and WordNet is used to determine sentiment score of each word. In this corpus, the word "police" has a positive sentiment score of 0.125 which is considered contradictory to the aims of this research. Therefore, in this study, the sentiment score for the word "police" was changed to 0 to avoid arising bias into the analysis conducted as the authors wanted to examine the sentiment of the Indonesian police. Furthermore, the corpus used is not equipped with attributes to perform similarity checks for each word to determine the most appropriate sentiment score to be used in a sentence, especially for words that have negating words such as "not", "no", and so on. To avoid errors in sentiment scoring for each word, a weighted average method is employed. This allows the determination of scores by taking the average of multiple sentiments present in a word.

However, the dataset in this case study was collected from a social media platform, Twitter, which is known for its informal nature, where users often use abbreviations, misspellings, and a mix of local languages that can produce errors and challenges in the analysis process. The dataset contains numerous spelling errors, as users on Twitter may not adhere to proper spelling conventions, which creates difficulties in accurately interpreting the sentiment of the text. Misspellings of words can lead to incorrect sentiment scores or even misinterpretations of the intended sentiment due to Twitter users mixing in local languages or dialects, which may not have direct translations or sentiment scores in the sentiment analysis corpus or Lexicon being used. This also challenges the accuracy of the sentiment determination of these specific expressions. Furthermore, the use of abbreviations is prevalent on Twitter, where users often shorten words or phrases for brevity that may not be accounted for in the sentiment analysis corpus or Lexicon, leading to inaccurate sentiment scores.

Similar challenges in the application of machine learning methods also emerged. They are caused by the strong relationship between feature sets and tokenized words. The tokenization process can be biased due to the vastness and complexity of the Indonesian language, which includes synonymous words, abbreviations, and regional languages. Tokenization is a crucial step in machine learning models as it breaks down each text into single tokens or words, which are then used to determine the features for further analysis. On the other hand, the case of Indonesian language processing has a wide range of synonyms, meaning that different words with similar meanings can be used interchangeably in different contexts, leading to variations in sentiment analysis results depending on which specific synonym is chosen as a token. Also, diverse Indonesian languages from different regions often have different meanings that might not be captured by the sentiment analysis model, affecting the accuracy of the sentiment analysis results. Therefore, it is important to accommodate comprehensive and contextually aware sentiment analysis models that can oversee the nuances of the Indonesian language as the actual dataset before the data preprocessing step becomes important to be developed properly.

To avoid the difficulties highlighted, researchers attempted to create a dictionary of abbreviated words with a conversion into standard Indonesian as shown at Table 4. On Lexicon Based analysis, this dictionary was implemented in the data preprocessing step before performing the text translation step. Converting abbreviations and slang words helps the machine translate the text well and improve its accuracy, as increasing translation accuracy can improve the whole analysis. This vocabulary was used before tokenization in machine learning analysis. The following stage is necessary for supplying the optimal token and feature set for the model. To further enhance performance, a more comprehensive dictionary covering a wide range of regional and daily languages should be developed.

An investigation of topic modeling aimed at examining keywords associated with the most frequently occurring primary topics within a collection of tweets accumulated over a specific time span was also done. The outcomes of this exploration are to determine specific issues that are concerning the society during the time span, as the results presented in Table 5. In the table, from 10 words to be determined, we found 8 words with weight value above 0. It is evident from the analysis that among the ten subtopics pertaining to the police domain, which were extracted using the LDA algorithm, a number of subtopics exhibit notable relevance to the overarching primary topic.

As the topic of this research is Indonesian police, the word police has a higher weight than the others. Thus, we can exclude the keyword police from the result and leave the rest of the words. On topic 1, both the words "jalan" and "kantor" stand out, and it can be concluded that topic 1 is related to street police and police station. Topics 2 and 7 have differences of 4 on the weight of the word "tau", and the additional word "nggak" on topic 7 represents the acknowledgement of police, which can be a positive or negative acknowledgement of police toward civilian occurrences. The word "lapor" appears in topic 4 along with the word "nggak" with a significant difference in weight, representing the topic of the behavior in reporting cases to police, which can be positive or negative. However, topics on 8 and 9 are related to "orang" and "kasih", which can respectively be concluded as wide and general topics, so we leave the analysis as it is.

Table 4. Example of dictionary for abbreviations and slang words

Abbreviation and Slang Words	Standard Indonesian	Translation
‘g’, ‘ga’, ‘gk’, ‘gx’, ‘gak’, ‘ngga’, ‘nggak’, ‘engga’, ‘enggak’	Tidak	Not
‘duit’, ‘cuan’, ‘dana’, ‘fulus’	Uang	Money
‘bgt’, ‘bet’, ‘banged’, ‘banget’	Sangat	Very

Table 5. Output of LDA algorithm

Words		Topic									
ID	EN	1	2	3	4	5	6	7	8	9	10
Polisi	Police	51.5	32.5	240	26.9	0.1	82.6	70.5	25.6	40.8	207.5
Nggak	Not	0.1	0.1	0.1	9.1	0.1	92.1	144.2	0.1	0.1	0.1
Kasih	Give/Affection	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	26.1	0.1
Jalan	Street	42.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Orang	Civilian	0.1	0.1	0.1	0.1	0.1	0.1	0.1	53.1	0.1	0.1
Tau	Know	0.0	11.6	0.1	0.1	0.0	0.0	15.6	0.1	0.1	0.1
Kantor	Office/Station	27.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Lapor	Report	0.1	0.0	0.1	36.1	0.1	0.1	0.1	0.1	0.0	0.1

5. CONCLUSION

The comparison of classifier algorithms for this case study shows that classification algorithms can be used for examining public trust sentiment against the police. Thus, the sentiment of Indonesian people represented by Twitter users toward POLRI is positive through analysis of several techniques, whose performances evaluated by F-score value and accuracy value of the model are greater than 50% with medium estimation of error, and the highest performance is held by Naive-Bayes algorithm. Performing sentiment analysis on Indonesian users of social media, particularly Twitter, can indeed be challenging due to various constraints such as diverse range of regional languages, the prevalence of abbreviations and slang words, and the difficulty in determining precise keywords for data crawling. The diverse range of regional languages in Indonesia poses a significant challenge as different regions may have their own unique expressions, idioms, and vocabulary and the model for sentiment analysis needs to account for regional variations in language usage and sentiment expression. Other than that, the abundance of abbreviations and slang words commonly used on social media platforms like Twitter further complicates sentiment analysis as informal expressions often have unclear representations in sentiment Lexicons or databases, making it difficult to accurately classify their sentiment polarity.

Moreover, determining appropriate and precise keywords for data crawling requires careful selection of keywords to obtain relevant and representative data. The choice of keywords influences the quality and relevance of the collected dataset, which directly impacts the accuracy and reliability of the sentiment analysis results. To overcome these challenges, researchers and developers need to employ comprehensive techniques that consider regional variations, account for informal language usage, and incorporate domain-specific Lexicons or custom sentiment dictionaries. Additionally, leveraging advanced natural language processing techniques, such as machine learning models trained on large-scale Twitter data, can help improve the accuracy and increase validity of sentiment analysis in the context of Indonesian social media data.

Apart from that, a topic modeling technique that uses previously chosen keywords as the main topic can be employed as well for grouping content from Twitter social media by topic. According to the study's findings, three sub-topic groups with the following information have been successfully modeled using the POLRI themed dataset:

(1) Police on the street or in a station
(2) Police involvement or acknowledgement in neighborhood events
(3) The action of contacting the police

It should be noted that the subtopics established through the model are based on datasets from a certain period of time and might not adequately represent concerns over a longer period of time. Given that public opinion toward the POLRI is dynamic and influenced by news or current occurrences, this model can be taken advantage of as a way of assessing and preserving public trust in the POLRI. Additionally, periodic evaluations on POLRI-related themes could potentially be investigated further in future studies to uncover patterns of concerns and occurrences in society that influence the dynamics of public sentiment towards POLRI.

In general, every method employed in this study is not limited to a single subject. The approach described here is also applicable for research on various kinds of issues because the abbreviation language library and dictionary used in this study are typical daily languages, resembling the language often used in Indonesian-based tweets.

ACKNOWLEDGMENT

This research receives no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Bayer, J.B., Triệu, P., Ellison, N.B. (2020). Social media elements, ecologies, and effects. Annual Review of Psychology, 71: 471-497. https://doi.org/10.1146/annurev-psych-010419-050944

[2] Danaditya, A., Ng, L.H.X., Carley, K.M. (2022). From curious hashtags to polarized effect: Profiling coordinated actions in Indonesian twitter discourse. Social Network Analysis and Mining, 12(1): 105. https://doi.org/10.1007/s13278-022-00936-2

[3] Robinson, P.H., Seaman, J., Sarahne, M. (2022). Standing back and standing down: Citizen non-cooperation and police non-intervention as causes of justice failures and crime. Hofstra Law Review, 51: 923.

[4] Siregar, S. (2019). Indonesian national police in terrorism handling policy during joko widodo's government: Analysis of role, function and evaluation. In International Conference on Democratisation in Southeast Asia (ICDeSA 2019). Atlantis Press, pp. 93-98. https://doi.org/10.2991/icdesa-19.2019.20

[5] Abdurrachman, H., Ari Sudewo, F. (2018). The use of violence in indonesian police investigation (cek similarity). International Journal of Engineering & Technology, 7(3.21): 497-501. https://doi.org/10.14419/ijet.v7i3.21.17221

[6] Zulganef, Z., Nilasari, I. (2022). Analysis of public trust and POLRI performance: An exploratory study. Publisia: Jurnal Ilmu Administrasi Publik, 7(2): 211-227. https://doi.org/10.26905/pjiap.v7i2.8247

[7] Tandel, S.S., Jamadar, A., Dudugu, S. (2019). A survey on text mining techniques. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, pp. 1022-1026. https://doi.org/10.1109/ICACCS.2019.8728547

[8] Shu, X., Ye, Y. (2023). Knowledge discovery: Methods from data mining and machine learning. Social Science Research, 110: 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

[9] Alsaeedi, A., Khan, M.Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 10(2): 361-374.

[10] Schröer, C., Kruse, F., Gómez, J.M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181: 526-534. https://doi.org/10.1016/j.procs.2021.01.199

[11] Schäfer, F., Zeiselmair, C., Becker, J., Otten, H. (2018). Synthesizing CRISP-DM and quality management: A data mining approach for production processes. In 2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD),

Marrakech, Morocco, pp. 190-195. https://doi.org/10.1109/ITMC.2018.8691266

[12] Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., Gurusamy, V. (2014). Preprocessing techniques for text mining. International Journal of Computer Science & Communication Networks, 5(1): 7-16.

[13] Kadhim, A.I. (2018). An evaluation of preprocessing techniques for text classification. International Journal of Computer Science and Information Security (IJCSIS), 16(6): 22-32.

[14] Pandey, S.V., Deorankar, A.V. (2019). A study of sentiment analysis task and it's challenges. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, pp. 1-5. https://doi.org/10.1109/ICECCT.2019.8869160

[15] Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10): 78-87. https://doi.org/10.1145/2347736.2347755

[16] Saravanan, R., Sujatha, P. (2018). A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 945-949. https://doi.org/10.1109/ICCONS.2018.8663155

[17] Rish, I. (2001). An empirical study of the Naive Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 3(22): 41-46.

[18] Xu, Y., Klein, B., Li, G., Gopaluni, B. (2023). Evaluation of logistic regression and support vector machine approaches for XRF based particle sorting for a copper ore. Minerals Engineering, 192: 108003. https://doi.org/10.1016/j.mineng.2023.108003

[19] Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20: 273-297. https://doi.org/10.1007/BF00994018

[20] Bhuta, S., Doshi, A., Doshi, U., Narvekar, M. (2014). A review of techniques for sentiment analysis of twitter data. In 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, pp. 583-591. https://doi.org/10.1109/ICICICT.2014.6781346

[21] Mowlaei, M.E., Abadeh, M.S., Keshavarz, H. (2018). A lexicon generation method for aspect-based opinion mining. In 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), Las Palmas de Gran Canaria, Spain, pp. 000107-000112. https://doi.org/10.1109/INES.2018.8523897

[22] Huang, Y., Wang, R., Huang, B., Wei, B., Zheng, S.L., Chen, M. (2021). Sentiment classification of crowdsourcing participants' reviews text based on LDA topic model. IEEE Access, 9: 108131-108143. https://doi.org/10.1109/ACCESS.2021.3101565

[23] Tolner, F., Takács, M., Eigner, G., Barta, B. (2021). Clustering of business organisations based on textual data-An LDA topic modeling approach. In 2021 IEEE 21st International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, pp. 000073-000078. https://doi.org/10.1109/CINTI53070.2021.9668337

[24] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools and Applications, 78: 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

[25] Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H. (2009). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer, 2: 1-758. https://doi.org/10.1007/978-0-387-84858-7

[26] Lavazza, L., Morasca, S. (2023). Common problems with the usage of F-measure and Accuracy metrics in medical research. IEEE Access, 11: 51515-51526, https://doi.org/10.1109/ACCESS.2023.3278996

[27] Sammut, C., Webb, G.I. (2017). Encyclopedia of machine learning and data mining. Springer Publishing Company, Incorporated.

[28] Guo, W. (2022). Applications of logistic regression and naive bayes in commodity sentiment analysis. In 2022 4th International Conference on Image, Video and Signal Processing, pp. 224-230. https://doi.org/10.1145/3531232.3531265

[29] Ilmania, A., Cahyawijaya, S., Purwarianti, A. (2018). Aspect detection and sentiment classification using deep neural network for Indonesian aspect-based sentiment analysis. In 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, pp. 62-67. https://doi.org/10.1109/IALP.2018.8629181

[30] Wu, X., Linghu, Y., Wang, T., Fan, Y. (2021). Sentiment analysis of weak-ruletext based on the combination of sentiment lexicon and neural network. In 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, pp. 205-209. https://doi.org/10.1109/ICCCBDA51879.2021.9442593

[31] Pan, D., Yuan, J., Li, L., Sheng, D. (2019). Deep neural network-based classification model for Sentiment Analysis. In 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), Beijing, China, pp. 1-4. https://doi.org/10.1109/BESC48373.2019.8963171

[32] Firmino Alves, A.L., Baptista, C.D.S., Firmino, A.A., Oliveira, M.G.D., Paiva, A.C.D. (2014). A comparison of SVM versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 FIFA confederations cup. In Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, pp. 123-130. https://doi.org/10.1145/2664551.2664561

[33] Genç, Ş., Surer, E. (2023). ClickbaitTR: Dataset for clickbait detection from Turkish news sites and social media with a comparative analysis via machine learning algorithms. Journal of Information Science, 49(2): 480-499. https://doi.org/10.1177/01655515211007746

[34] Latif, S., Shafait, F., Latif, R. (2021). Analyzing LDA and NMF topic models for urdu tweets via automatic labeling. IEEE Access, 9: 127531-127547. https://doi.org/10.1109/ACCESS.2021.3112620

[35] Moussa, R., Sarro, F. (2022). On the use of evaluation measures for defect prediction studies. In Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 101-113. https://doi.org/10.1145/3533767.3534405