Semantic Feature Engineering with LSA-SVM for Cyberbullying Comment Classification on Instagram

Wulandari, Haerunnisya Makmur, Dewi Fatmarani Surianto*, Andi Akram Nur Risal, Nur Azizah Eka Budiarti, Satria Gunawan Zain, Abdul Wahid
Department of Computer Engineering, State University of Makassar, Makassar, Indonesia
E-mail: wlndry.nurdin@gmail.com, haerunnisyamakmur44@gmail.com, dewifatmaranis@unm.ac.id, andi.akram@unm.ac.id, nurazizah.ekabudiarti@gmail.com, satria.gunawan.zain@unm.ac.id, wahid@unm.ac.id
*Corresponding author

Keywords: classification, comments, cyberbullying, instagram, latent semantic analysis, support vector machine.

Received: August 25, 2024

Social media is now an essential part of everyday life, with Instagram being one of the most popular platforms and often utilized for various purposes, one of which is to increase popularity. However, the platform also often becomes a place where acts of violence and impoliteness in commenting increase, known as cyberbullying. To address the problem, detecting and classifying cyberbullying comments on Instagram is an important step in cyberbullying prevention. However, in text classification, several challenges need to be overcome to ensure the success of the model, such as polysemy, curse of dimensionality, and selection of text representation for feature extraction. Therefore, this study aims to implement a feature engineering technique using a hybrid approach that combines word weighting with TF-IDF and LSA method to reduce feature dimensionality and capture the semantic meaning of the data, with SVM used as a classifier to classify bullying and non-bullying comments. The results showed that the proposed method using feature engineering of the LSA matrix formed from the dataset of one of the classes, yielded a significant accuracy of 97%. In comparison, the conventional method with feature engineering using TF-IDF and the use of LSA matrix formed from the dataset of both classes only achieved an accuracy of 84%. This shows that the proposed method is more effective than the baseline approach.

Povzetek: Študija raziskuje klasifikacijo kibernetskega ustrahovanja v komentarjih na Instagramu z uporabo semantičnega inženiringa značilk s hibridnim pristopom LSA-SVM. Predlagana metoda združuje TF-IDF za uteževanje besed in LSA za zmanjšanje dimenzionalnosti značilk in zajemanje semantičnega pomena. Uporaba SVM kot klasifikatorja je pokazala, da ta pristop dosega dobro učinkovitost pri odkrivanju in klasifikaciji komentarjev kibernetskega ustrahovanja.

1 Introduction

Nowadays, digital technologies such as mobile device and social media are not just additional amenities, but have become an essential part of the daily lives of global citizens. More than 66% of the global population uses the internet, with active social media users reaching 5.04 billion by the start of 2024, representing a 5.6% increase in the past year [1]. Indonesia is one of the countries with the largest number of social media users, reaching 139 million users or equivalent to 49.9% of the total population [2]. Instagram has become one of the most popular social media platforms, with around 16.5% of internet users between the ages of 16 and 64 choosing Instagram over other platforms [1]. In Indonesia itself, there are around 106 million active users on the platform [3]. This phenomenon illustrates how social media, especially Instagram, has become an integral part of the daily lives of Indonesians.

Instagram as social media can be utilized to form an online community and share information, ideas, personal messages, and other content [4]. Users of this platform also utilize it for various purposes such as earning income as an endorser, improving existence, self-image, and popularity by sharing various types of content, in the hope of getting attention from other users through symbol responses, comments, or simply viewing [5]. However, this platform often becomes a place where acts of violence and incivility in commenting are on the rise [6]. Negative comments and hostile private messages are part of cyberbullying [7]. Based on a survey involving more than 10,000 young people aged 12 to 20 years old shows that cyberbullying is widespread, with nearly 70% of teens admitting to perpetrating abusive behavior towards others online and 17% claiming to have been victims of online bullying [8]. These cyberbullying behaviors can cause physical or psychological harm to their victims, including stress, social isolation, low self-esteem, anxiety, and depression [9] [10].

Table 1: Research components

No.	Component	Details
1	Technologies	Natural Language Processing (NLP), Machine Learning, Classification Models, and Cyberbullying Detection
2	Tools	Python, Scikit-learn and Jupyter Notebook
3	Algorithms	Support Vector Machine (SVM), TF-IDF, Latent Semantic Analysis (LSA), and Confusion Matrix
4	Case Studies	Cyberbullying detection from Instagram comments
5	Datasets	Instagram comment datasets related to cyberbullying and non-cyberbullying in the Indonesian language
6	Methods	Machine learning using the SVM algorithm for comment classification, feature extraction using a combination of TF-IDF and LSA, and evaluation using a confusion matrix

Cyberbullying on Instagram has become a common problem with serious consequences for individuals' mental health, therefore detecting and classifying cyberbullying comments is an important step in preventing the spread of this harmful behavior early on [11]. Machine learning-based classification models can be used to detect cyberbullying, as they have been proven efficient in predicting and detecting various types of data, including text data in the form of comments [12]. However, text classification faces several challenges such as word polysemy [12] [13], high data dimensionality that triggers overfitting [15], and text representations that affect the model's ability to understand text meaning [16]. Thus, the effectiveness of a classification model depends on the feature extraction results used, so the discovery of active feature extraction techniques has been the focus of many researchers to improve text classification performance [15]. Some of the simplest and most commonly used text representations for feature extraction are Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) [11] [17]. Although they can represent text well, these approaches tend to produce text representations that have large dimensions and limitations, such as BoW's inability to account for word order and TF-IDF's lack of semantic context [12].

Therefore, this study aims to identify comments containing elements of cyberbullying on social media, particularly on Instagram, by developing a machine learning-based classification model. To classify texts using machine learning, feature engineering techniques using hybrid approaches such as a combination of TF-IDF word weighting method and Latent Semantic Analysis (LSA) method are proposed to reduce the dimensionality of the features, while capturing the semantic meaning of the data. This research also investigates the effect of multiple feature engineering performed on classification performance, using features on the whole data as well as on one of the classes only. In summary, the main contribution of this research is the introduction of a feature engineering approach to improve the performance of the model, so that it can properly distinguish between cyberbullying and non-cyberbullying comments. The effectiveness of the proposed feature engineering technique is evaluated using confusion matrix using the Support Vector Machine algorithm as the classifier. This is important as it provides new insights into the problem of cyberbullying comments on social media and proposes a new method to address it. Table 1 summarizes the key components of this research, including the technologies, tools, algorithms, case studies, datasets, and methods used to achieve the research objectives.

2 Related works and novelty of the proposed work

Some previous research shows that there are various techniques developed to classify reviews or comments using machine learning algorithms. Several techniques and types of features have been used, including the use of the TF-IDF model as a feature for classification. One of the studies related to sentiment analysis on Shopee app reviews used TF-IDF as feature extraction, and Support Vector Machine (SVM) and Random Forest (RF) were used as classifiers. The results showed that the SVM model had a higher accuracy of 84.71% compared to Random Forest which was 82.21% [18]. A similar study used TF-IDF as feature extraction and Naive Bayes (NB) as a classifier for sentiment analysis of game products at Shopee, with an accuracy of 80.22% [19].

In addition, some studies focus on semantic modeling in the text as a feature extraction scheme. Several studies proposed semantic methods using LSA to improve model performance for detecting adverse drug reactions with four machine learning algorithms used as classifiers including SVM, NB, Logistic Regression (LR), and Artificial Neural Network (ANN) with two document representations used, namely Term Frequency (TF) and TF-IDF. The results showed that LSA as a feature with LR and ANN classifiers outperformed other algorithms with an accuracy of 82% [20] [21] [22]. Other studies used NB, SVM, and LR methods combined with LSA for sentiment analysis of tweet replies on public figure accounts, with the highest accuracy on LR of 80.6% [23].

Various studies have also been conducted to identify and detect cyberbullying in recent years. One such study used Natural Language Processing (NLP) techniques and machine learning algorithms to detect cyberbullying in Bangla and Bangla Romanization texts from YouTube comments, where the SVM method achieved 76% accuracy for the Bangla dataset, while Multinomial Naive Bayes (MNB) achieved 84% accuracy for the Bangla Romanization dataset and 80% for the combined dataset [24]. Another study performed cyberbullying classification from Twitter data using SVM method as a classifier and Information Gain (IG) as a feature selection technique, by exploring the effect of various SVM parameters and various IG selection thresholds [25]. There is also research applying machine learning techniques using three datasets to detect cyberbullying, where SVM achieved the highest accuracy of 92% [26].

Furthermore, research that proposed an approach to detect cyberbullying in Roman Urdu texts by addressing the colloquial and non-standard variations of users' writing styles on social media, using several feature extraction techniques such as N-Gram, hybrid n-gram, and TFIDF weighting. Experimental results showed that SVM with hybrid N-gram embedded features achieved the highest average accuracy of about 83% [27]. In the following, cyberbullying identification using SVM, LR, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM, and Bidirectional Encoder Representations from Transformers (BERT) methods. The results showed that BERT achieved the highest F1-score, which was 94% for the Twitter dataset, 91% for the Wikipedia dataset, and 92% for the Formspring dataset [28].

The next research is the classification of cyberbullying comments from tweets on Twitter using the ANN method where the classification results were improved using Deep Reinforcement Learning (DRL), resulting in an average increase in classification accuracy of 80.69% [29]. Another study was to detect cyberbullying in tweets from Twitter using TF-IDF feature extraction and Naive Bayes and SVM algorithms for the classification process, where the accuracy of Naive Bayes was 52.70% while SVM reached 71.25% [30].

Table 2: Related works and proposed method

Ref. (Year)	Domain (Language)	Class (Data Source)	Method (Accuracy)	Limitations/Descriptions
[17] (2023)	Shopee App Reviews (Indonesian)	Positive and negative (Kaggle)	TF-IDF+SVM (82.21%) TF-IDF+RF (84.71%)	Limited to Bag-of-Words and TF-IDF features; lacks semantic understanding or contextual embeddings.
[18] (2021)	Game product reviews (Indonesian)	Positive, neutral, negative (Shoppe)	TF-IDF+Naive Bayes (80,22%)	Did not compare Naive Bayes with other advanced models or integrate semantic/contextual embeddings.
[19][20] (2021)	Adverse Drug Reactions (English)	Positive negative (Previous research)	TF+LSA+LR (82%) TF-IDF+LSA+SVM (80%)	Did not utilize advanced techniques like word embeddings or neural networks, limiting performance in complex semantic tasks.
[21] (2024)	Adverse Drug Reactions (English)	Positive and negative (Previous research)	TF+ANN (82%) TF+LSA+ANN (85%) TF-IDF+LSA+ANN (83%)	Did not employ advanced embedding techniques or architectures like transformers, which could enhance contextual understanding.
[22] (2022)	Public Figure (Indonesian)	Positive, neutral, negative (Twitter)	TF-IDF+LSA+NB (78.6%) TF-IDF+LSA+LR (80.6%) TFIDF+LSA+SVM (80.4%)	The addition of LSA reduced model performance; lacked exploration of hybrid or ensemble methods for improvement.
[23] (2021)	Cyberbullying (Bangla and Romanized Bangla)	Bullying and not-bullying (Youtube)	TF-IDF+SVM (76%) - Bangla TF-IDF+MNB (84%) - Romanized TF-IDF+MNB (80%) - Bangla and Romanized	Relied solely on TF-IDF without addressing deeper contextual relationships or semantic nuances between words.

Ref. (Year)	Domain (Language)	Class (Data Source)	Method (Accuracy)	Limitations/Descriptions
[24] (2020)	Cyberbullying (Indonesian)	Bullying and not-bullying (Twitter)	TF-IDF+SVM (75%) TF-IDF+IG+SVM (76.66%)	Dependency on TF-IDF and IG limited the feature set, potentially missing critical semantic patterns.
[25] (2022)	Cyberbullying (English)	Bullying and non-bullying. (Previous research)	Random Forest (91%) Naïve Bayes (87%) SVM (92%)	Did not fully explore deep contextual embeddings or advanced neural architectures for nuanced text understanding.
[26] (2023)	Cyberbullying (Roman Urdu)	Bullying and non-bullying. (Social media)	TF-IDF+Hybrid N-gram+SVM (83%)	High-dimensional feature space due to N-gram combination; lacked deeper semantic feature extraction.
[27] (2022)	Cyberbullying (English)	Positive negative (Twitter, Wikipedia, Formspring)	BERT (94%) - Twitter BERT (91%) - Wikipedia BERT (92%) - Formspring	Fine-tuning on the BERT model requires extensive training time and significant computational resources.
[28] (2021)	Cyberbullying (English)	Bullying, and non-bullying (Twitter)	Deep Reinforcement Learning (80.69%)	Integrating ANN with DRL for improved classification. However, it adds complexity in implementation and processing time, limiting scalability.
[29] (2020)	Cyberbullying (English)	Bullying and non-bullying. (Twitter)	TF-IDF+NB (52.70%) TF-IDF+SVM (71.25%)	Relatively low accuracy; lacked advanced feature engineering and semantic understanding.
Proposed Method	**Cyberbullying (Indonesian)**	**Bullying and non-bullying. (Instagram)**	**(Feature Engineering TF-IDF + LSA) + SVM (97%)**	**Outperforms previous methods in accuracy and balance**

As explained in the previous section, many techniques have been applied by researchers to solve classification problems. The various techniques are analyzed and compared based on their performance and the type of dataset used. Table 2 presents a summary of various machine learning approaches and techniques applied, as well as the proposed method.

Several previous studies have shown that NLP approaches can be applied to detect and classify texts, especially in the context of cyberbullying. However, most of these studies focus on the English language or employ single approaches, such as TF-IDF or standard machine learning algorithms, which often fail to capture the complex semantic meanings in textual data. Additionally, high-dimensional feature representations frequently lead to overfitting issues. Therefore, the novelty of this study, as summarized in Table 2, lies in proposing a hybrid approach that integrates TF-IDF and LSA to capture semantic context and reduce feature dimensionality while exploring the impact of feature engineering techniques on the performance of SVM-based classification models. This research makes a significant contribution to the development of cyberbullying classification models in the Indonesian language, an area that remains underexplored. Evaluation results demonstrate that the proposed approach is more effective than the baseline, achieving high performance in detecting comments containing cyberbullying.

3 Methodology

This section explains the various stages involved in completing the research on cyberbullying comment classification using the proposed method. These stages include data collection, data annotation, preprocessing, feature engineering, classification model, and evaluation. The scheme of the stages of this research can be seen in Figure 1.

3.1 Data collection

The dataset used was obtained from several sources with a total of 2100 data shown in Table 3. The data used in this study were Instagram comments taken from the posts of

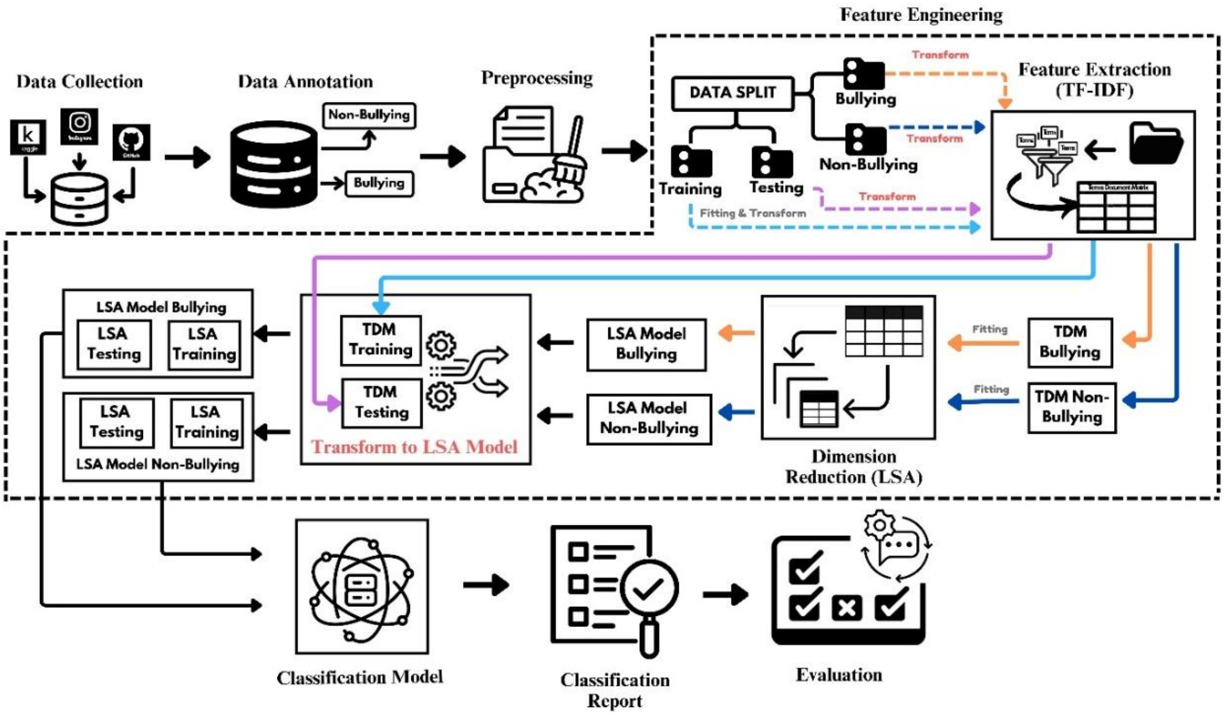

Figure 1: Stages of the proposed research method

several artists in Indonesia, as cyberbullying which includes negative comments, personal attacks, and ridicule, is a common problem that often targets celebrities and influencers [31]. This data collection technique was carried out to obtain a dataset that is representative of the phenomenon of cyberbullying on social media, particularly within the social context of Indonesia. This relevant data supports the training of the model to more accurately identify and classify bullying comments.

Table 3: Dataset distribution

Source	Number of Comments	
	Bullying	Non-Bullying
Instagram	525	525
Kaggle [32]	325	325
GitHub [33]	200	200
Total	1050	1050

3.2 Data annotation

Annotation on the dataset aims to provide information related to the category or class by each data, where in this study it consists of two classes, namely bullying and non-bullying comments. Determination of annotation was done by analyzing the comments based on their characteristics as in Table 4 [34]. This process is crucial to ensure accurate data labeling, enabling the machine learning model to better recognize specific patterns in each category, thereby supporting more reliable classification.

Table 4: Characteristics of bullying and non-bullying comments

No.	Bullying	Non-Bullying
1.	Contains insults or harassment	Contains support or appreciation
2.	Disrespectful or contains abusive language	Not condescending and not scornful

3.3 Preprocessing

Preprocessing is done to clean and improve the structure of the comment text so that it is more easily processed by the algorithm or model used [35] [36]. This stage is the most important initial step in classification, where the combination of preprocessing techniques can affect the classification performance results [37]. By cleaning the text of irrelevant elements enables the model to concentrate more effectively on essential information required for classification. The preprocessing steps used in the research include casefolding, regex, stopword removal, and stemming. The data preprocessing stages can be seen in Figure 2.

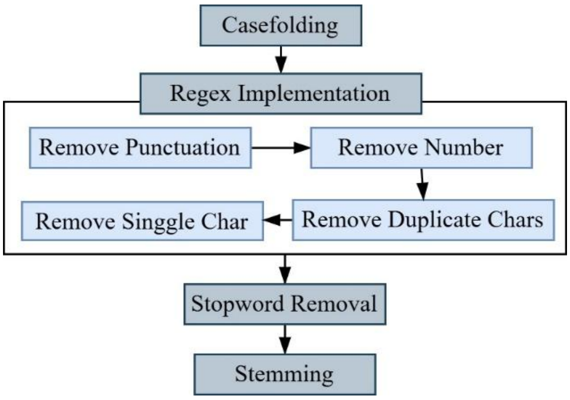

Figure 2: Data preprocessing stages

In this stage of text preprocessing, several libraries available in the Python programming language were used. First, to perform casefolding, the lower () function was used to convert all letters in the text to lowercase. Next, the regular expressions (regex) library was used to implement text processing based on the specified pattern. As for removing common words (stopwords) in Indonesian, the Natural Language Toolkit (NLTK) library was used by retrieving the list of available stopwords. To perform stemming (removal of prefixes and suffixes) in Indonesian, the Sastrawi library developed specifically for this language was used.

3.4 Feature engineering

Feature engineering is the process of extracting meaning from raw data by converting text into numerical values, which aims to improve efficiency and consistency in text classification using machine learning approach [38]. This technique is necessary to capture patterns and characteristics of relevant text data, enabling the model to better understand the relationships between words. In this study, the feature engineering technique used involves two main methods, namely feature extraction using TF-IDF and dimensionality reduction with LSA. The algorithm or steps of the feature engineering process along with their descriptions can be found in Table 5 below.

Table 5: Algorithm of the feature engineering

Algorithm: Feature Engineering for Text Classification using TF-IDF and LSA	
Step 1: Data Preparation	
1.	**Input:** Preprocessed dataset
2.	**Process:** • Categorize text data into bullying and non-bullying groups to facilitate the extraction of more specific features. • Split the dataset into training and testing subsets with a 70:30 ratio using train_test_split to ensure the model is trained on the majority of the data (70%) and tested on the remaining data (30%), allowing for an objective evaluation of the model's performance.
3.	**Output:** Training and testing datasets for both bullying and non-bullying categories
Step 2: Feature Extraction using TF-IDF	
1.	**Input:** Training and testing text data
2.	**Action:** • Use TfidfVectorizer to convert text into numerical representations • Training Phase: Apply fit_transform () on the training set to generate Term-Document Matrix (TDM). • Testing Phase: Apply transform () on the test set using the trained TF-IDF model for consistency • Bullying and Non-Bullying Datasets: Perform the same transformation process on the bullying and non-bullying datasets to create feature representations specific to each category and enhance the model's ability to recognize patterns within each category.
3.	**Output:** TDM from the training set, testing set, bullying dataset, and non-bullying dataset.
Step 3: Dimensionality Reduction using LSA	
1.	**Input:** TDM from TF-IDF for the training and testing sets, as well as the bullying and non-bullying datasets.
2.	**Action:** • Apply TruncatedSVD to reduce the dimensionality while preserving important information in the data. • **Fit the LSA model:** Apply the LSA model to the bullying and non-bullying TDM datasets using fit() with 500 topics on different TruncatedSVD models to capture relevant patterns and topics from bullying or non-bullying data. This will make the model more effective in identifying differences between the two and generating better dimensional representations. • **Transform Data:** Use the trained LSA model, both with the bullying and non-bullying datasets, to transform the TDM of the training and testing sets into lower-dimensional representations through the U matrix from Singular Value Decomposition (SVD). This matrix is then used as features in the classification model to improve the efficiency and accuracy of predictions.
3.	**Output:** • Matrix U on the training set and testing set using the LSA model trained with the Bullying TDM.

Algorithm: Feature Engineering for Text Classification using TF-IDF and LSA	
	• Matrix U on the training set and testing set using the LSA model trained with the Non-bullying TDM.

3.4.1 Feature extraction

Feature extraction using the Term Frequency-Inverse Document Frequency (TF-IDF) method aims to transform text into a numerical representation in the form of a Term-Document Matrix (TDM), allowing the model to understand the data quantitatively. The feature extraction process began by splitting the dataset into two separate subsets, one for training the model (training set) and one for testing the model (testing set), with a proportion of 70:30 using the `train_test_split` function from the Scikit-learn library. In addition, the dataset was further categorized into bullying and non-bullying datasets, for more specific feature extraction. This division helped to understand and identify the unique characteristics of each comment type and improved the model's performance in detecting and classifying cyberbullying more effectively.

The TF-IDF method was used to give weight to words in a document based on their frequency of occurrence, both in the document itself and in the entire corpus, thus enabling the identification of more meaningful and relevant words. The mathematical equations for calculating the weight of words in a document using TF-IDF can be found in equations (1), (2), and (3) as follows [39].

$$tf_{t,d} = \begin{cases} 1 + log_{10}\, tf_{t,d}, & if\ tf_{t,d} > 0 \\ 0 & otherwise \end{cases} \quad (1)$$

$$idf_t = log_{10} \frac{N}{df_t} \quad (2)$$

$$w_{t,d} = tf_{t,d} \times idf_t \quad (3)$$

Description:
$tf_{t,d}$ = frequency of occurrence of word t in document d
N = number of documents
idf_t = number of documents that contain t
$w_{t,d}$ = TF-IDF weight

In this research, the feature extraction process used `TfidfVectorizer` from the Scikit-learn library to transform text data into TF-IDF representation in TDM form. The feature extraction process in this study involved fitting and transforming the data using the TfidfVectorizer with fit_transform () function on the training set. After that, the transform () function of TfidfVectorizer was applied using the testing set, to transform the data into the same TDM format as learned from the training set. A similar transformation process was also applied to the bullying dataset and the non-bullying dataset.

3.4.2 Dimensionality reduction

Dimensionality reduction was performed using the Latent Semantic Analysis (LSA) method, which can extract and represent the hidden meaning of documents in a text corpus by reducing the dimensionality of the data. The application of LSA at this stage aims to enhance the model's understanding of the semantic context of the text by reducing dimensional complexity, allowing the model to capture deeper relationships between words, even if those words do not frequently appear together in the same document. This method utilized Singular Value Decomposition (SVD) to reduce the number of dimensions of the TDM used as LSA input. This process produced three new matrices from the SVD decomposition, with the mathematical equation that can be seen in equation (4) below [40].

$$A_{m,n} = U_{m,m} \Sigma_{m,n}\ (V^T)_{n,n} \quad (4)$$

Description:
$A_{m,n}$= m × n matrix (m documents, n terms)
$U_{m,m}$ = m × r matrix (m documents, r concepts)
$\Sigma_{m,n}$ = r × r matrix (value of each concept)
$(V^T)_{n,n}$ = n × r matrix (n terms, r concepts)

In this research, the LSA method was implemented using `TruncatedSVD` from the scikit-learn decomposition library. The implementation began with fitting using the fit() function on the TDM generated from the `TfidfVectorizer` transformation on the bullying dataset and non-bullying dataset. As a result, two LSA models were obtained, namely bullying and non-bullying LSA models, with 500 topics defined for comment classification. These models were then used to transform the TDM training set and testing set into the LSA model. From the results of the LSA model transformation, the matrix (U), which represented the relationship between documents and latent concepts in the dataset, then was used as a feature for the classification process.

3.5 Classification model

Classification models were developed using the Support Vector Machine (SVM) algorithm, which can be used in machine learning for classification. The purpose of applying the SVM algorithm is to leverage SVM's ability to handle complex and non-linearly separable data, with the Radial Basis Function (RBF) kernel chosen for its capability to capture non-linear relationships in the data. [41]. The SVM model was implemented using the scikit-learn library by utilizing the `SVC` class. The classification model development process involved exploring various feature engineering scenarios to identify the most optimal features used in the formation of the classification model. The feature engineering scenarios explored in the classification model building can be found in Table 6 below.

Table 6: Feature engineering scenarios

Scenario	Classification feature formation			
	TF-IDF		LSA	
	Bullying training set	Non-bullying training set	Bullying dataset	Non-Bulling dataset
1	✓	✓	-	-
2	✓	-	-	-
3	-	✓	-	-
4	✓	✓	✓	✓
5	✓	✓	✓	-
6	✓	✓	-	✓

From Table 6 above, there are six scenarios performed to form classification features based on the subset of data used. The aim of each scenario is to generate features representing the bullying class, the non-bullying class, or both. The results of the feature formation from each scenario will be used to transform or extract features from the training and testing data so that the data can be used to train and test the classification model.

3.6 Evaluation

The performance measurement of the model is based on data from the confusion matrix, which aims to provide a comprehensive overview of how the model classifies comments into bullying and non-bullying categories. By calculating accuracy, precision, recall, and F1-Score, the evaluation is conducted on various aspects of the model's quality, including its ability to correctly identify bullying comments (precision), recognize all existing bullying comments (recall), and balance between the two (F1-Score). The use of this confusion matrix allows for a deeper evaluation of the model's strengths and weaknesses, providing clearer insights into areas that need improvement. The form of the confusion matrix is presented in Table 7.

Table 7: Confusion matrix

		Prediction	
		TRUE	FALSE
Actual	TRUE	True Positive	False Negative
	FALSE	False Positive	True Negative

The formulas for calculating accuracy, precision, recall, and F1-Score can be seen in equations (5), (6), (7), and (8) respectively [42].

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1\text{-}Score = \frac{2 \times precision \times recall}{precision + recall} \quad (8)$$

Table 8: Preprocessing results

No.	Preprocessing Stage	Bullying	Non-Bullying
1	Original data	Laki skrng kan pada begitu bnyk nya😂 seramm yaa.. Modal tampang aja..eeh numpang hidup sama istri😄 lebih baik menjanda😌	Dari kasus yg seperti ini.. Bahwa Sampai kapanpun.. Rezeki Halal itu udah yg paling Bener.. Meski Tak Banyak.. Namun Berkah 🙏🖤
2	Casefolding	laki skrng kan pada begitu bnyk nya😂 seramm yaa.. modal tampang aja..eeh numpang hidup sama istri😄 lebih baik menjanda😌	dari kasus yg seperti ini.. bahwa sampai kapanpun.. rezeki halal itu udah yg paling bener.. meski tak banyak.. namun berkah 🙏🖤
3	Regex	laki skrng kan pada begitu bnyk nya seram ya modal tampang aja eh numpang hidup sama istri lebih baik menjanda	dari kasus yg seperti ini bahwa sampai kapanpun rezeki halal itu udah yg paling bener meski tak banyak namun berkah
4	Stopword removal	laki skrng bnyk seram modal tampang aja eh numpang hidup istri menjanda	rezeki halal udah bener berkah
5	Stemming	laki skrng bnyk seram modal tampang aja eh numpang hidup istri janda	rezeki halal udah bener berkah

4 Results and discussion

Data pre-processing is an important step as raw data obtained from various sources is often not in a form that is ready for use. Therefore, the preprocessing stage is necessary to obtain a more structured dataset to produce informative features. The following is an example of the results of the preprocessing stage in Table 8.

In the casefolding stage, the letters in the text were all converted into lowercase letters. Then, the regex stage went through several stages of the process, namely the removal of punctuation marks, numbers, double characters, and single characters. In the example above, it can be seen that all full stops in the text are removed and the word "seramm" is changed to "seram". Then, common words that often appear in the text were removed in the stopword removal stage from the NLTK corpus such as the words "pada", "sama", "lebih", "dari", "paling", and others, resulting in a shorter sentence than before. In the stemming stage, the word was converted to its base word form such as "menjanda" was converted to "janda". As for the non-bullying example sentence, there was no word change because all words were already in their base word form.

After preprocessing, the dataset was divided into two data partitions for classification model building. A total of 70% of the 2100 data was allocated as training data, while the rest became testing data. In Figure 3, there is an even distribution in each data partition, with the amount of data divided proportionally for each class without significant differences.

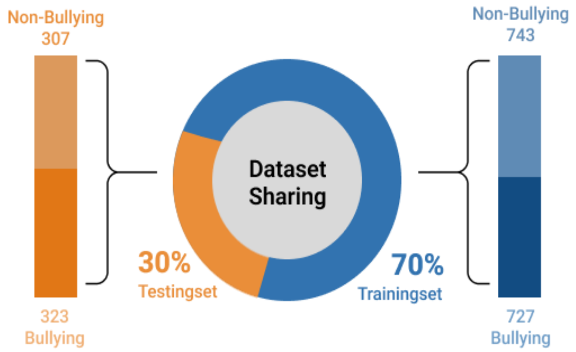

Figure 3: Distribution of dataset sharing

In developing the classification model, a features engineering stage was carried out which involved feature extraction using the TF-IDF and LSA methods. By using both feature extraction methods, several feature engineering scenarios were carried out for the formation of classification features to produce an optimal classification model. In Table 9 below are the classification model performance results based on the features engineering scenarios in Table 6.

Table 9: Results of research scenario classification

Scenario	Accuracy (%)	Precision (%)		Recall (%)		F1-Score	
		Bullying	Non- Bullying	Bullying	Non- Bullying	Bullying	Non- Bullying
Scenario 1	84	82	87	89	79	85	83
Scenario 2	84	81	89	91	77	86	83
Scenario 3	82	83	81	81	82	82	81
Scenario 4	84	80	90	92	76	86	82
Scenario 5	97	95	100	100	94	97	97
Scenario 6	97	100	95	95	100	97	97

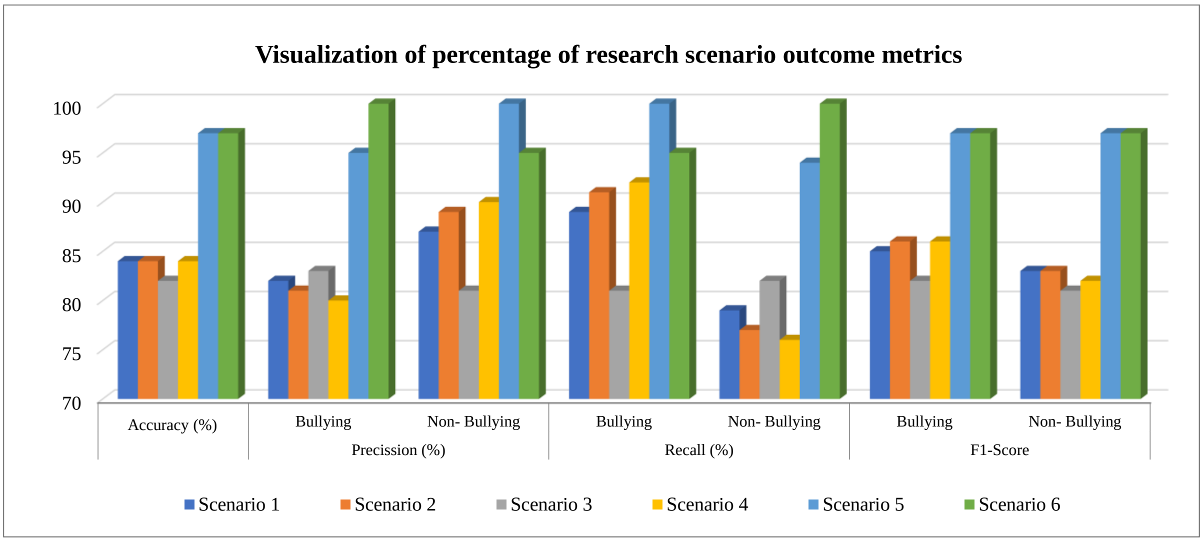

Figure 4: Research scenario results diagram

Based on the results obtained in Table 9, in scenario 1, the classification features were generated from weighting the documents using the TF-IDF matrix trained with the entire train data. Although the classification model had good precision in identifying the Bullying and Non-Bullying categories, there was a significant difference in the recall rate between them. In scenarios 2 and 3, the training process only used one of the classes to form the classification features based on the TF-IDF matrix. However, the results showed that in scenario 2 where the classification features were formed from the TF-IDF of the bullying class, there was a decrease in the non-Bullying class compared to the previous scenario. Otherwise, in scenario 3, the classification features obtained from the TF-IDF matrix of the non-bullying class showed a stable and consistent performance of the model in classifying texts for both classes, but a significant decrease in performance compared to scenarios 2 and 3.

The results of scenarios 1, 2, and 3 still showed low performance in distinguishing bullying and non-bullying classes. This was due to the limitation of the model that used TF-IDF as feature extraction which has not been able to handle synonyms, polysemy, and hidden meanings in the text as LSA does. [43]. Therefore, another features engineering scenario applying the LSA method was required, such as in scenarios 4, 5, and 6. In those scenarios, the LSA method was implemented after word weighting using the TF-IDF matrix to obtain the U matrix from SVD result which was used as a classification feature. The TDM used as LSA input in this scenario was the TF-IDF matrix trained using the entire train data. In scenario 4, the classification features were obtained from the LSA method trained using the TDM from the training set, but no performance improvement occurred.

Based on the diagram in Figure 4, scenario 5 which was the proposed feature engineering technique using the LSA method trained using TDM from the bullying class dataset, showed excellent performance with accuracy, precision, recall, and F1-Score values reaching 100% for both categories. Scenario 6 which was also a feature engineering technique proposed using TDM from the non-bullying class dataset for LSA modeling, yielded almost identical performance to scenario 5. Although Scenarios 5 and 6 had similar performance, Scenario 5 had slightly lower recall for the non-Bullying category, while Scenario 6 had slightly lower precision and F1-Score for the Bullying category. Therefore, it can be concluded that the proposed method of LSA model building using data from only one of the classes showed excellent accuracy results when used for classification model training.

Although LSA was expected to improve accuracy by capturing semantic meaning, Figure 4 shows that in scenario 4, the result is almost lower than other TF-IDF scenarios. This was because the data used had a phenomenon of lexical ambiguity where there were 25% of the same terms appeared in both classes, but the terms had different meanings. The terms that appeared together in two classes can be seen in Figures 5 and 6 below.

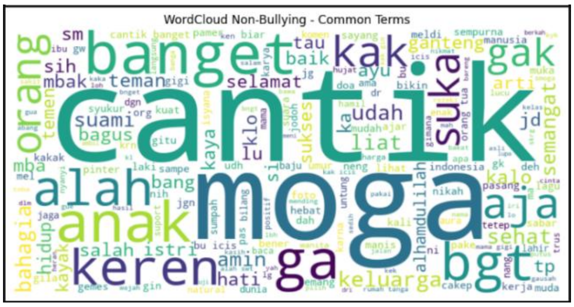

Figure 5: Wordcloud non-bullying terms found in bullying class

Figure 6: Wordcloud bullying terms found in non-bullying class

Based on Figures 5 and 6, the word "cantik" which means physically attractive or having a beautiful appearance is usually used in a positive context. In Figure 5, it can be seen that the word "cantik" appears frequently in non-bullying class. However, this word also appears in the bullying class, as shown in Figure 6, but the frequency is not too much. Another example is the word "ganteng" which also has a positive meaning in the non-bullying class, but this word is also found in the bullying class. Based on this, the application of the LSA method in both classes caused the phenomenon of lexical ambiguity, which is the same word, but has different meanings in different contexts [44]. Lexical ambiguity resulted in mixed semantic information which made it difficult for the model to learn the true meaning of the word "cantik" or "ganteng" when used in the context of satire (bullying) and a positive context (non-bullying). The phenomenon of mixed semantic information can also be seen from the topics generated in scenario 6 with the LSA model formed using the two classes dataset presented in Table 10 below.

Table 10: Example terms on topics in scenarios 4, 5, 6

Scenario Topic 4	**Scenario Topic 5**	**Scenario Topic 6**
[‘suka’, ‘sih’, ‘bgt’, ‘udah’, ‘hidup’, ‘anjing’, ‘penjara’, ‘sehat’, ‘cakep’, ‘sarah’]	[‘ganteng’, ‘doang’, ‘ga’, ‘muka’, ‘udah’, ‘anjing’, ‘jelek’, ‘gak’, ‘cari’, ‘modal’]	[‘cantik’, ‘banget’, ‘masya’, ‘alah’, ‘moga’, ‘kak’, ‘bgt’, ‘keren’, ‘sehat’, ‘icis’]

Based on Table 10, in scenario 4, it can be seen that there are positive terms such as ‘suka’, ‘hidup’, ‘sehat’ and ‘cakep’, which are quite representative of the non-bullying class. But within the same topic, there are also some blasphemous words such as 'anjing' and 'penjara' which represent the bullying class. This created confusion of meaning for the model, as terms that usually represent one class are mixed into one topic. When these conflicting terms appeared together, the model struggled to determine the true context and sentiment of the text. For example, the word 'suka' usually carries a positive connotation, associated with compliments or harmless comments. Otherwise, the word 'anjing' in this context is a slur with strong negative connotations, often used to insult or demean someone. This combination of conflicting terms in one topic caused the model to struggle to classify the text correctly, as the terms it received were ambiguous and conflicting. This situation is an example of mixed semantic information, where the text contains various conflicting elements of meaning.

As for the topic of Scenario 5 found in Table 10, the terms generated almost all mean bullying. For example, words like 'anjing' and 'jelek' strongly reflect bullying content. Although there is the word 'ganteng' which connotatively can be categorized as non-bullying terms, the number was very small or only appeared a few times. With such a strong dominance of bullying terms in this topic, the LSA model did not experience any confusion in capturing the semantic meaning of the word, so the LSA model was able to clearly identify that the overall context of the topic was more inclined towards bullying despite some exceptions.

Therefore, experimenting with training LSA in a single class showed effective results as the model could learn the meaning of words more effectively without experiencing confusion from different meanings in other contexts. [45]. This allowed the LSA to better capture the special characteristics of the class compared to using two classes, due to noise reduction i.e. noise due to lexical ambiguity. Thus, Scenario 5 and Scenario 6 showed very high model performance with accuracy and F1-Score reaching 97% for both Bullying and Non-Bullying categories. The scenario results showed that the classifier was better able to recognize patterns from the data when the classification features from the LSA were sourced from one class, even though each used an LSA implementation for classification feature formation. The confusion matrix of the two scenarios can be seen in Tables 11 and 12.

Table 11: Confusion matrix of scenario 5 classification results

		Prediction	
		Bullying	Non-Bullying
Actual	Bullying	323	0
	Non-Bullying	18	289

Table 12: Confusion matrix of scenario 6 classification results

		Prediction	
		Bullying	Non-Bullying
Actual	Bullying	306	17
	Non-Bullying	1	306

From Table 11, it can be seen that in scenario 5, the classification model tended to classify non-bullying as bullying, where there are 17 non-bullying comments predicted as bullying. Meanwhile, in Table 12, the classification model shows the opposite tendency, classifying bullying comments as non-bullying. In this scenario, there are 17 bullying comments predicted as non-bullying by the model. Despite the good performance, the phenomenon shown in Table 11 and Table 12 indicates a significant challenge in the context-based text classification process. Although the model had been trained to recognize certain patterns, be it bullying or non-bullying, the tendency of the classification model to predict the class was highly dependent on the data used during the training of the LSA model.

Table 13: Performance Comparison Results with Available Approaches

Approaches and Method		Accuracy (%)	Precision (%)		Recall (%)		F1-Score	
			Bullying	Non-Bu	Bullying	Non-Bull	Bullying	Non-Bull
TFIDF + IG + SVM		83	82	85	86	80	84	82
TFIDF + Chi-Square + SVM		83	81	86	88	78	84	82
TFIDF + PCA + SVM		84	80	91	93	75	86	82
TFIDF + Ngram + SVM	Bigram	60	75	56	34	88	47	68
	Trigram	54	53	76	98	8	69	15
Proposed Method	Scenario 5	97	95	100	100	94	97	97
	Scenario 6	97	100	95	95	100	97	97

Table 13 presents a comprehensive performance evaluation of our proposed method against various established methodologies. The comparison considers accuracy, precision, recall, and F1-score across the "Bullying" and "Non-Bullying" classes.

The table reveals that our proposed method (Scenarios 5 and 6) consistently achieves the highest performance metrics, with an accuracy of 97% and F1-scores of 97% for both classes. Compared to traditional approaches like TFIDF + IG + SVM and TFIDF + PCA + SVM, which exhibit accuracies of 83% and 84%, respectively, our method demonstrates a significant improvement. Moreover, while bigram and trigram-based models perform poorly, especially in recall for the "Non-Bullying" class (8% for trigrams), our method excels with a recall of 95% for "Bullying" and 100% for "Non-Bullying."

This comparison substantiates the efficacy of our approach, showcasing its superiority in accurately identifying and classifying bullying behavior. The results also underline the robustness of the proposed method in achieving balanced precision and recall, which is critical for practical applications.

5 Conclusion and future work

Based on the research conducted, it was found that the formation of classification features from several scenarios resulted in significant variations in accuracy. Training the model from the TF-IDF matrix formed from the entire dataset that included bullying and non-bullying texts resulted in an accuracy of 84%. The accuracy obtained when the model was only trained using a subset of the dataset containing bullying text only resulted in the same accuracy of 84%, while the model trained using a subset of the non-bullying dataset resulted in a slightly lower accuracy of 82%. Furthermore, using the LSA matrix of the entire dataset showed that the accuracy remained at 84%, which was the same as using TF-IDF on the entire dataset. The phenomenon of lexical ambiguity was the main cause of this non-optimal accuracy. Lexical ambiguity occurs when the same word appears in both classes (bullying and non-bullying), but has different meanings. This made it difficult for the model to accurately learn the meaning of the word in the right context. To solve the lexical ambiguity problem, the proposed method was to form the LSA matrix from only one of the dataset classes. When LSA was applied to only a subset of bullying or non-bullying datasets, the resulting accuracy increased significantly to 97%. This improvement showed that the model could more effectively classify texts into bullying or non-bullying because the LSA was able to better capture the specific characteristics of one class.

For future development, it should be noted that the tendency of the classification model to predict the class is highly dependent on the data used during the training of the LSA model. In addition, this model only achieved maximum accuracy for negative or positive sentiments. Therefore, further development is recommended for creating a multi-class classification model that can recognize neutral, negative, and positive sentiments.

Variables and constants used

To ensure a comprehensive and systematic approach, this study incorporates several variables and constants, as outlined in Table 14 below.

Table 14: Nomenclature of variables and constants

Symbol	Descriptions
t	Term, a word or phrase that appears in a collection of documents, analyzed in the calculation of TF (Term Frequency) and IDF (Inverse Document Frequency)
d	Document, a unit of text in a collection of documents analyzed for term frequency and term weighting
N	The total number of documents in the collection
$tf_{t,d}$	The frequency of term t in document d
idf_t	Inverse Document Frequency for term t, a measure that indicates how frequently term t appears in the document collection.
$w_{t,d}$	The weight of term t in document d, calculated as the product of TF and IDF
m	The number of rows in the matrix during matrix decomposition, such as in Singular Value Decomposition (SVD)
n	The number of columns in the matrix during matrix decomposition, such as in SVD
r	The number of principal components used in dimensionality reduction or LSA matrix decomposition
A	The resulting matrix from decomposition used in SVD or LSA
U	The orthogonal matrix of singular vectors for columns in SVD decomposition
Σ	The diagonal matrix of singular values
V	The orthogonal matrix of singular vectors for rows in SVD decomposition
TP	True Positive, the number of predictions classified as positive that are truly positive based on the actual data
TN	True Negative, the number of predictions classified as negative that are truly negative based on the actual data
FP	False Positive, the number of predictions classified as positive but are actually negative based on the actual data
FN	False Negative, the number of predictions classified as negative but are actually positive based on the actual data

References

[1] Datareportal, 'Digital 2024: Global Overview Report', DataReportal – Global Digital Insights. Accessed: Mar. 27, 2024. [Online]. Available: https://datareportal.com/reports/digital-2024-global-overview-report

[2] Datareportal, 'Digital 2024: Indonesia', DataReportal – Global Digital Insights. Accessed: Mar. 27, 2024. [Online]. Available: https://datareportal.com/reports/digital-2024-indonesia

[3] Datareportal, 'Instagram Users, Stats, Data, Trends, and More', DataReportal – Global Digital Insights. Accessed: Mar. 27, 2024. [Online]. Available: https://datareportal.com/essential-instagram-stats

[4] H. W. Aripradono, 'Penerapan Komunikasi Digital Storytelling Pada Media Sosial Instagram', *Teknika*, vol. 9, no. 2, pp. 121–128, Nov. 2020, doi: https://doi.org/10.34148/teknika.v9i2.298

[5] R. Rubiyanto and M. Fildyanti, 'Personal Branding Barbie Kumalasari Untuk Meraih Popularitas Melalui Instagram', *WACANA J. Ilm. Ilmu Komun.*, vol. 20, no. 1, Jun. 2021, doi: https://doi.org/10.32509/wacana.v20i1.1253

[6] M. A. Caesaryo, M. Giswandhani, and A. Z. Hilmi, 'Cyberbullying Selebriti Instagram', *J. Syntax Admiration*, vol. 3, no. 5, pp. 671–679, May 2022, doi: https://doi.org/10.46799/jsa.v3i5.423

[7] C. Juditha, 'Analysis of Content the Case of Cyberbullying Against Celebrities on Instagram', *J. Penelit. Komun. Dan Opini Publik*, vol. 25, no. 2, 2021, doi: 10.33299/jpkop.25.2.4300.

[8] J. Wakefield, 'Instagram tops cyber-bullying study', Jul. 18, 2017. Accessed: Mar. 27, 2024. [Online]. Available: https://www.bbc.com/news/technology-40643904

[9] M. S. Z. Al-Sulami, 'The Role of Social Work in Facing the Negative Effects of Cyberbullying on Adolescents in Saudi Arabia', *Arab J. Sci. Res. Publ.*, vol. 7, no. 11, pp. 109–124, Nov. 2023, doi: https://doi.org/10.26389/ajsrp.n130723

[10] Kus Hanna Rahmi, Rijal Abdillah, and Andreas Corsini Widya Nugraha, 'Understanding The Danger of Bullying: A Phenomenological Study on Female College Students As Victims of Cyberbullying', *Krtha Bhayangkara*, vol. 18, no. 1, pp. 61–84, Apr. 2024, doi: https://doi.org/10.31599/krtha.v18i1.1612

[11] D. Samalo, R. Martin, and D. N. Utama, 'Improved Model for Identifying the Cyberbullying based on Tweets of Twitter', *Informatica*, vol. 47, no. 6, Jun. 2023, doi: https://doi.org/10.31449/inf.v47i6.4534

[12] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, 'Text Classification Algorithms: A Survey', *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: https://doi.org/10.3390/info10040150

[13] T. Gupta and E. Kumar, 'Learning Improved Class Vector for Multi-Class Question Type Classification', presented at the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021), Bangalore, India, 2021. doi: https://doi.org/10.2991/ahis.k.210913.015

[14] F. Di Martino and S. Senatore, 'Semi-supervised Feature Selection Method for Fuzzy Clustering of Emotional States from Social Streams Messages', in *Advances in Machine Learning/Deep Learning-based Technologies*, vol. 23, G. A. Tsihrintzis, M. Virvou, and L. C. Jain, Eds., in Learning and Analytics in Intelligent Systems, vol. 23. , Cham: Springer International Publishing, 2022, pp. 9–25. doi: https://doi.org/10.1007/978-3-030-76794-5_2

[15] A. Adeleke, N. A. Samsudin, Z. A. Othman, and S. K. Ahmad Khalid, 'A two-step feature selection method for quranic text classification', *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 2, p. 730, Nov. 2019, doi: https://doi.org/10.11591/ijeecs.v16.i2.pp730-736

[16] D. Kim, 'Research On Text Classification Based On Deep Neural Network', *Int. J. Commun. Netw. Inf. Secur. IJCNIS*, vol. 14, no. 1s, pp. 100–113, Dec. 2022, doi: https://doi.org/10.17762/ijcnis.v14i1s.5618

[17] V. Dogra *et al.*, 'A Complete Process of Text Classification System Using State-of-the-Art NLP Models', *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–26, Jun. 2022, doi: https://doi.org/10.1155/2022/1883698

[18] S. Suswadi and Moh. Erkamim, 'Sentiment Analysis of Shopee App Reviews Using Random Forest and Support Vector Machine', *Ilk. J. Ilm.*, vol. 15, no. 3, pp. 427–435, Dec. 2023, doi: https://doi.org/10.33096/ilkom.v15i3.1610.427-435

[19] R. Kosasih and A. Alberto, 'Sentiment analysis of game product on shopee using the TF-IDF method and naive bayes classifier', *Ilk. J. Ilm.*, vol. 13, no. 2, pp. 101–109, Aug. 2021, doi: https://doi.org/10.33096/ilkom.v13i2.721.101-109

[20] A. A. Nafea, N. Omar, and M. M. AL-Ani, 'Adverse Drug Reaction Detection Using Latent Semantic Analysis', *J. Comput. Sci.*, vol. 17, no. 10, pp. 960–970, Oct. 2021, doi: https://doi.org/10.3844/jcssp.2021.960.970

[21] A. H. Abed, S. A. Jabber, and A. A.-J. Altameemi, 'Extracting Adverse Drug Reaction Using Latent Semantic Analysis from Medical Social Media Reviews', 2021, *ICIC International 学会*: 08. doi: 10.24507/icicel.15.08.907.

[22] A. A. Nafea, N. Omar, and Z. M. Al-qfail, 'Artificial Neural Network and Latent Semantic Analysis for Adverse Drug Reaction Detection', *Baghdad Sci. J.*, May 2023, doi: https://doi.org/10.21123/bsj.2023.7988

[23] M. A. Gumilang, T. D. Puspitasari, H. A. Putranto, A. Kholiq, and A. Samsudin, 'Sentiment Analysis Based on Tweet Reply at Public Figure Account using Machine Learning and Latent Semantic Analysis', in *2022 8th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia: IEEE, Sep. 2022, pp. 1–6. doi: https://doi.org/10.1109/icst56971.2022.10136288

[24] Md. T. Ahmed, M. Rahman, S. Nur, A. Z. M. T. Islam, and D. Das, 'Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts', *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 20, no. 1, p. 89, Feb. 2021, doi: https://doi.org/10.12928/telkomnika.v20i1.18630

[25] N. M. G. D. Purnamasari, M. A. Fauzi, I. Indriati, and L. S. Dewi, ‘Cyberbullying identification in twitter using support vector machine and information gain based feature selection’, *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, no. 3, p. 1494, Jun. 2020, doi: https://doi.org/10.11591/ijeecs.v18.i3.pp1494-1500

[26] A. Ali and A. M. Syed, ‘Cyberbullying Detection using Machine Learning’, *Pak. J. Eng. Technol.*, vol. 3, no. 2, pp. 45–50, Apr. 2022, doi: https://doi.org/10.51846/vol3iss2pp45-50

[27] A. Dewani *et al.*, ‘Detection of Cyberbullying Patterns in Low Resource Colloquial Roman Urdu Microtext using Natural Language Processing, Machine Learning, and Ensemble Techniques’, *Appl. Sci.*, vol. 13, no. 4, p. 2062, Feb. 2023, doi: https://doi.org/10.3390/app13042062

[28] S. Paul and S. Saha, ‘CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification’, *Multimed. Syst.*, vol. 28, no. 6, pp. 1897–1904, Dec. 2022, doi: https://doi.org/10.1007/s00530-020-00710-4

[29] N. Yuvaraj *et al.*, ‘Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking’, *Math. Probl. Eng.*, vol. 2021, pp. 1–12, Feb. 2021, doi: https://doi.org/10.1155/2021/6644652

[30] R. R. Dalvi, S. Baliram Chavan, and A. Halbe, ‘Detecting A Twitter Cyberbullying Using Machine Learning’, in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India: IEEE, May 2020, pp. 297–301. doi: https://doi.org/10.1109/iciccs48265.2020.9120893

[31] A. R. Lahitani, A. N. Zhafarina, N. S. Windi Oktavia, and N. Jariyah, ‘Pemetaan Topik Pembicaraan Pada Komentar Live Youtube Menggunakan K-Means Clustering sebagai Identifikasi awal Kejahatan Verbal Cyberbullying’, *J. Tek. Elektro Uniba JTE UNIBA*, vol. 8, no. 2, pp. 399–403, Apr. 2024, doi: https://doi.org/10.36277/jteuniba.v8i2.253

[32] C. T. Hanni, ‘Cyberbullying Bahasa Indonesia’. Accessed: Jul. 05, 2024. [Online]. Available: https://www.kaggle.com/datasets/cttrhnn/cyberbullying-bahasa-indonesia

[33] R. S. Perdana, ‘Dataset Sentimen Analisis Bahasa Indonesia’, GitHub. Accessed: Jul. 05, 2024. [Online]. Available: https://github.com/rizalespe/Dataset-Sentimen-Analisis-Bahasa-Indonesia/blob/master/dataset_komentar_instagram_cyberbullying.csv

[34] M. Jubaidi and N. Fadilla, ‘Pengaruh Fenomena Cyberbullying Sebagai Cyber-Crime di Instagram dan Dampak Negatifnya’, *Shaut Al-Maktabah J. Perpust. Arsip Dan Dok.*, vol. 12, no. 2, pp. 117–134, Dec. 2020, doi: https://doi.org/10.37108/shaut.v12i2.327

[35] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, ‘Big data preprocessing: methods and prospects’, *Big Data Anal.*, vol. 1, no. 1, p. 9, Dec. 2016, doi: https://doi.org/10.1186/s41044-016-0014-0

[36] H. Makmur, W. Wulandari, D. F. Surianto, and M. Fajar B, ‘Analisis Sentimen Penghapusan Skripsi sebagai Tugas Akhir Mahasiswa Menggunakan Metode Multi-Layer Perceptron’, *Komputika J. Sist. Komput.*, vol. 13, no. 2, Oct. 2024, doi: https://doi.org/10.34010/komputika.v13i2.12402

[37] S. Khairunnisa, A. Adiwijaya, and S. A. Faraby, ‘Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)’, *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 2, p. 406, Apr. 2021, doi: https://doi.org/10.30865/mib.v5i2.2835

[38] A. N. Sutranggono, Riyanarto Sarno, and Imam Ghozali, ‘Multi-Class Multi-Level Classification of Mental Health Disorders Based on Textual Data from Social Media’, *J. Inf. Commun. Technol.*, vol. 23, no. 1, pp. 77–104, Jan. 2024, doi: https://doi.org/10.32890/jict2024.23.1.4

[39] R. Dzisevic and D. Sesok, ‘Text Classification using Different Feature Extraction Approaches’, in *2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Vilnius, Lithuania: IEEE, Apr. 2019, pp. 1–4. doi: https://doi.org/10.1109/estream.2019.8732167

[40] D. Kalman, ‘A Singularly Valuable Decomposition: The SVD of a Matrix’, *Coll. Math. J.*, vol. 27, no. 1, pp. 2–23, Jan. 1996, doi: https://doi.org/10.1080/07468342.1996.11973744

[41] R. Jeevitha, K. Chaitanya, N. Mathesh, B. Nithyanarayanan, and P. Darshan, ‘Using Machine Learning to Identify Instances of Cyberbullying on Social Media’, in *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode, India: IEEE, Mar. 2023, pp. 207–212. doi: https://doi.org/10.1109/icscds56580.2023.10104720

[42] M. Grandini, E. Bagli, and G. Visani, ‘Metrics for Multi-Class Classification: an Overview’, *ArXiv*, vol. abs/2008.05756, Aug. 2020, doi: https://doi.org/10.48550/arXiv.2008.05756

[43] S. Qaiser and R. Ali, ‘Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents’, *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: https://doi.org/10.5120/ijca2018917395

[44] F. Zait and N. Zarour, ‘Addressing Lexical and Semantic Ambiguity in Natural Language Requirements’, in *2018 Fifth International Symposium on Innovation in Information and Communication Technology (ISIICT)*, Amman: IEEE, Oct. 2018, pp. 1–7. doi: https://doi.org/10.1109/isiict.2018.8613726

[45] S.-A. Rueschemeyer and M. G. Gaskell, Eds., *The Oxford Handbook of Psycholinguistics*, 2nd ed. Oxford University Press, 2018. doi: https://doi.org/10.1093/oxfordhb/9780198786825.001.0001