Sistemasi: Jurnal Sistem Informasi
Volume 13, Nomor 4, 2024: 1587-1596
ISSN:2302-8149
e-ISSN:2540-9719

Sentiment Analysis of Cyberbullying Detection on Social Networks using the Sentistrenght Method

[1]Kevin Heryadi Yunior, [2]Anik Vega Vitianingsih*, [3]Slamet Kacung, [4]Anastasia Lidya Maukar, [5]Andini dwi arumsari
[1,2,3]Informatics Department, Universitas Dr.Soetomo, Surabaya, Indonesia
[4]Industrial Engineering Department, President University, Bekasi, Indonesia
[5]Department of Psychology, Universitas Muhammadiyah Surabaya, Indonesia
*email: *vega@unitomo.ac.id*

(***received*: 7 June 2024, *revised:* 19 June 2024, *accepted:* 20 June 2024**)

Abstract

In today's swiftly changing digital realm, social media has emerged as a pervasive means of communication, yet it has also fostered the rise of cyberbullying, especially among young demographics. This research strives to develop an application that assesses public sentiment on Instagram regarding cyberbullying instances, categorizing sentiments as positive, negative, or neutral. Drawing data from Instagram accounts such as *kumparandotcom*, *merdekadotcom*, and *okezonedotcom*, the approach combines lexicon-based text labelling and sentiment analysis employing Sentistrength. Findings demonstrate the method's effectiveness, achieving accuracy, precision, and recall rates exceeding 85% while offering precise visualization of predictions. This study contributes to combatting cyberbullying, aiming to improve victims' mental well-being by providing clearer insights into social sentiment. The dataset comprises 4500 comments collected through web crawling, categorized into positive (735 entries), negative (2478 entries), and neutral (1288 entries) sentiments. The evaluation highlights the commendable performance of Sentistrength, achieving the highest accuracy at 93.85%.

Keywords: cyberbullying, sentiment analysis, social media, Instagram, sentistrength

1 Introduction

The rapid development of the digital era has turned social media into one of the most popular communication tools accessible to anyone worldwide. Cyberbullying, defined as acts of violence experienced by children or adolescents from peers through the internet or cyberspace, has become a prevalent phenomenon [1]. It occurs when children or teenagers are attacked, mocked, intimidated, or embarrassed by their peers through various digital media, the internet, or mobile phones. Victims of cyberbullying are often children or teenagers who are subjected to ridicule and humiliation related to their physical appearance, skin colour, family background, or behaviour in the school environment [1]. With the advancement of technology, the use of social media, especially Instagram, has significantly increased. This phenomenon expands the opportunities for negative behaviours like cyberbullying, as these platforms facilitate digital interaction. Children and teenagers are more vulnerable to cyberbullying due to their active involvement in the online world, adding complexity to efforts to address and protect them [2].

The progress of modernity, innovation, and widespread dissemination of social media has led to an increase in the reach and intensity of negative behaviours such as cyberbullying. Digital bullying critically impacts the mental well-being of victims, with some facing suicide risks due to difficulties in managing the pressures they face [3]. Cyberbullying perpetrators also experience consequences, particularly prolonged feelings of guilt. Meanwhile, cyberbullying victims generally experience feelings of hurt and disappointment as the main consequences of the treatment they receive. Thus, both perpetrators and victims of cyberbullying suffer negative psychological impacts, creating a complex dynamic in the social media ecosystem [4].

Previous literature studies on public opinions regarding cyberbullying on social media have utilized methods such as K-Nearest Neighbour with hashtag #cyberbullying, achieving an accuracy rate of 71.43% [5]. On the other hand, another study employing hashtags #KuliahDaring, #sentistrength,

 ISSN:2302-8149
e-ISSN:2540-9719

and #klarifikasi, along with the Sentistrength method, achieved an accuracy rate of 85% [6]. This indicates that the Sentistrength method outperforms the K-NN method in accuracy, averaging 85% [6]. However, there is no research yet utilizing the Lexicon-based method to identify sentiment for labelling by assigning specific weights to comments and Sentistrength for classifying sentiments from labelling results, using Instagram datasets to generate netizen opinions on cyberbullying in a related case study.

This research aims to develop an application capable of sentiment analysis on netizen comments regarding cyberbullying and bullying cases on Instagram, classifying them into positive, negative, and neutral sentiments related to bullying cases. The dataset originates from Instagram accounts *kumparandotcom*, *merdekadotcom*, and *okezonedotcom*, covering various related cases. In contrast, the lexicon-based method is used for text labelling, and Sentistrength is employed to analyze sentiment levels [7]. The novelty of this research lies in utilizing a broader and varied dataset with accuracy, precision, and recall rates above 85%, as well as the addition of TF-IDF calculations, faster processing, and visualization in prediction and actual diagrams [7].

2 Literature Review

Research on sentiment analysis in social media has developed rapidly in recent years, with public sentiment analysis being applied in various studies. One frequently used method is SentiStrength, which is applied in various social media platform contexts. A recent study in the Journal of Social Science (JSS) titled "Detection of Cyberbullying on Facebook Using K-Nearest Neighbor Algorithm" demonstrates the ability of the K-Nearest Neighbor (KNN) algorithm to classify and detect cyberbullying with high accuracy [5]. This research emphasizes the effectiveness of KNN in identifying negative or insulting comments on Facebook, enabling users to take necessary preventive actions. However, the study found that computational time was a constraint, especially during testing with lemma normalization.

Meanwhile, the research [6] discusses the application of SentiStrength in sentiment analysis on Twitter. This research notes accuracy rates of up to 85% in classifying sentiments related to online learning [6]. The strength of this study lies in its ability to collect and process tweet data effectively, including data pre-processing and sentiment classification. However, the journal also notes an error rate of about 15%, indicating room for further improvement. These findings provide valuable insights into public responses to online learning policies and highlight areas needing further development to enhance sentiment analysis accuracy [6]. The research [7] contributes significantly to understanding sentiment analysis related to cyberbullying on the social media platform Twitter using the Support Vector Machine (SVM) method. This research demonstrates that SVM can transform and classify tweet texts with an accuracy rate of 70% [7]. However, the study emphasizes the need to increase the amount of training data to achieve higher results. The journal also highlights the importance of sentiment analysis in addressing cyberbullying issues on social media and identifying the need for further development in the methods used. These findings are highly relevant in understanding and addressing cyberbullying, providing a strong basis for future research [7].

Despite various studies demonstrating the effectiveness of methods such as KNN and SVM in sentiment analysis, there is still a gap in understanding how linguistic factors and cultural contexts influence sentiment analysis results, especially on platforms like Instagram. This research focuses on exploring this under-researched area using the SentiStrength method. SentiStrength is known for accurately measuring the strength of positive and negative sentiments in short texts. However, its implementation on Instagram, which has specific language use and visual context characteristics, has rarely been studied. Therefore, this research not only expands our understanding of user sentiment on Instagram but also provides new contributions that can be applied to address unresolved issues in existing literature. This study aims to fill the gap in the literature by providing new contributions to understanding sentiment analysis in social media, considering diverse linguistic and cultural factors. With a more comprehensive approach, it is hoped that the results of this research can provide more accurate and applicable recommendations in a broader context.

 ISSN:2302-8149
e-ISSN:2540-9719

3 Research Method

Throughout this study, Instagram comments data related to a specific topic were collected from January 2021 to December 2023. The data analysis process from web crawling data collection was conducted from January 2022 to June 2023. During this period, researchers implemented steps according to the research plan that had been previously formulated. Figure 1 shows the procedure used in this research experiment.

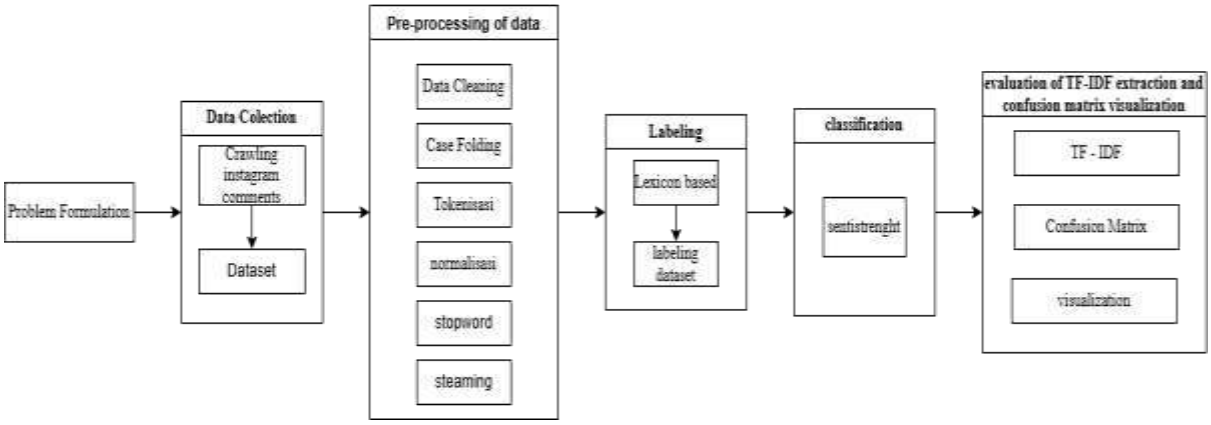

Figure 1. Research procedures

Source of Data Collection

Instagram comment data was searched and retrieved using the Instagram API with the keyword "cyberbullying" and the hashtags #cyberbullying, #child bullying, and #bullying. The system asks users to log in using their Instagram username and password to access and collect relevant comments. It should be noted that the user's login information will not be stored in the system. The collected comments are then manually filtered. Only those that are in Indonesian and do not contain images will be saved in CSV or XLSX file formats.

Data Collection Procedures

Retrieval of potential cyberbullying comment data from social media platforms is done through a crawling process using modules provided by the Python library. The successfully retrieved data is then stored as a raw dataset. This dataset will undergo several additional processing stages before it can be used for further analysis to identify and effectively address cyberbullying issues.

Pre-processing of data

In the data preparation stage, the raw dataset undergoes important processes. The first step is data cleaning, which aims to remove data that is irrelevant or contains errors to guarantee data quality. Next, case folding is performed to equalize the font format to keep it consistent. The next process is tokenization, where the text is broken down into smaller tokens to facilitate analysis. After that, normalization is applied to homogenize the data format. Then, stopwords or insignificant common words are removed. Finally, a stemming process returns the word to its base form, facilitating further data processing and analysis [8]. The following are the data preparation stages:

a) Data Cleaning: In this process, irrelevant, duplicate, or problematic data can be detected and removed, improving the quality and accuracy of the dataset used. By cleaning the data carefully, the analysis results will be more reliable and provide more valuable information.

b) Case Folding: At this stage, all letters in the text are adjusted to lowercase to ensure consistency in the analysis, reducing the complexity that may arise due to irrelevant case variations.

c) Tokenisasi: The tokenization stage is a crucial step that divides the text into smaller parts, such as words or phrases. This process supports further text analysis, facilitating an understanding of the sentence structure and the recognition of patterns and signification contained therein.

 ISSN:2302-8149
e-ISSN:2540-9719

d) Normalization: Word normalization in the pre-processing stage is a step to equalize various forms of words that have similar meanings into a standard form. For example, all words can be converted to lowercase, and plural words can be converted to singular. The goal is to allow text analysis to be performed consistently without being affected by variations in the form of the same word.
e) Stopword: stopwords refer to common words that appear frequently but have little contribution to text analysis. Stopword removal is done to cleanse the text of less relevant elements so that attention can be focused on more important keywords in data analysis.
f) Steaming: Steaming involves returning words in a text to their basic form, making analysis easier and reducing data duplication. It allows words related to the same root word to be grouped, increasing the effectiveness and precision of text analysis.

Identifying Data Categories

Each entity or text is analyzed to determine its sentiment in the data category identification process. Frequently used approaches are using sentistrength and lexicon-based methods. These methods categorize text into sentiment categories such as positive, negative, or neutral [8]. For example, a range from -5 to -1 indicates negative sentiment with different levels of strength, while a range from 1 to 5 indicates positive sentiment. Data that does not show a strong sentiment is labelled as neutral with a value of 0. This process helps understand the opinions or views in the text in more detail, providing a better understanding of analysis and decision-making in Table 1.

Table 1. Emotion data from Instagram comments with the keyword 'cyber bullying

emotions	Comment
-3 negative	*Besok besok nasdem pkb gabung perintah anies tinggal deh*
-4 negative	*jakarta doang gampang kibuli*
0 neutral	*Hehehehehehe*
5 positive	*masya allah*
3 positive	*beliau konsisten*

Sentistrenght

SentiStrength is a sentiment mining program developed by CyberEmotion [9]. SentiStrength is an example of a classifier that uses a lexicon-based approach to detect sentiment strength [10]. SentiStrength uses a two-scale vocabulary, making it possible to show that people can experience positive, negative, and neutral emotions up to a certain threshold [11]. SentiStrength produces both positive and negative results [10]. The range of values/scores is like the numbers 1 to 5. A value of 1 indicates that the data has no positive or negative feelings, and a value of 5 indicates that the data has negative or positive feelings. Only the maximum number of each emotion is recorded as the final number. This final score is still assigned additional values of 1 (positive) and -1 (negative) to prevent the sentiment from going to 0 or empty. Despite strong emotions, the situation remains [10]. SentiStrength can be modified to support different languages by extending the default database and meeting the requirements for specific language structures. SentiStrength features can be changed or altered to meet language requirements. Different language applications also result in lower accuracy values, as the analysis and development process is more often focused on English [11]. If a text has more positive than negative words, then the text data will be labelled with a positive sentiment. Vice versa, but if a sentence has the same number of positive and negative sentences, it will be labelled neutral. The process is described in Equation (1).

if positive value > negative value ; positive sentiment
if positive value < negative value ; negative sentiment (1)
if positive value = negative value ; neutral sentiment

 ISSN:2302-8149
e-ISSN:2540-9719

Assessment of Data Precision

Before the confusion matrix formation process, the first step is calculating TF-IDF to prepare the data. Sentiment classification testing on the comment data set is done by comparing the predicted results of the SentiStrength algorithm with the actual data automatically labelled into positive, negative, and neutral categories using a lexicon-based method. The SentiStrength algorithm does not require training data because it operates unguided, so testing is done by dividing the test data. The confusion matrix is formed in a 3x3 model with predicted and actual data. The confusion matrix Table can be seen in Table 2.

Table 2. Confusion matrix

		PREDICTION DATA		
		POSITIVE	NEGATIVE	NEUTRAL
ACTUAL DATA	Positive	TP	FN	FN
	Negative	FP	TP	TN
	Neutral	FP	TN	TP

Based on the confusion matrix calculation, the performance of a classification model can be measured as follows [12].

a) TP (True Positive) refers to the amount of data classified as positive and truly positive according to the actual data.

b) FP (False Positive) refers to the amount of data classified as negative, but the actual data is positive.

c) TN (True Negative) refers to the amount of data classified as negative and negative according to the actual data.

d) FN (False Negative) refers to the amount of data classified as positive, but the actual data is negative.

The calculation formula from the explanation above can be seen in equations 2, 3, and 4.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}} \text{ x } 100\% \quad (4)$$

4. Results and Discussions

In this research, Sentistrength and Sastrawi's lexicon-based methods are used along with the NLTK library to perform data pre-processing. This combination of tools provides a comprehensive framework for managing data. This method allows researchers to optimize data quality before performing sentiment analysis [13]. Data collection of cyberbullying comments on social media is carried out using the crawling method using the module provided by the Python library. The dataset that has been obtained is a raw dataset in Table 3.

Table 3. Crawling results from Instagram comments

No	Comment
1	*Paling enak pura pura gak tahu, terkaget kaget biar kelihatan dimedsos bikin satgas, selanjutnya EGP*
2	*Itu efek main game online jd niru tingkah game online* 😔
3	*Percuma gak bakal di penjara*
4	*Kepsek dam gurunya donk skalian*

 ISSN:2302-8149
e-ISSN:2540-9719

After the data collection was completed, processing could not begin immediately due to significant disruptions in the dataset. Therefore, the pre-processing stage became crucial for data cleaning, aiming to remove irrelevant attributes in the subsequent steps. The process can be seen in Table 4.

Table 4. Pre-processing results of data from comments

Data Pre-Processing	Result
Dataset	*Itu efek main game online jd niru tingkah game online* 😔
Data Cleaning	*Itu efek main game online jd niru tingkah game online*
Case Folding	*itu efek main game online jd niru tingkah game online*
Tokenisasi	*"itu" "efek" "main" "game" "online" "jd" "niru" "tingkah" "game" "online"*
Normalization	*"itu" "efek" "main" "game" "online" "jadi" "niru" "tingkah" "game" "online"*
Stopword	*"efek" "main" "game" "online" "niru" "tingkah" "game" "online"*
Steaming	*"efek" "main" "game" "online" "niru" "tingkah" "game" "online"*

In machine learning, data labelling involves assigning informative labels to raw data such as images, text, or videos. It aims to provide the context for machine learning models to learn and produce more precise predictions or results [14]. For example, labels can indicate the presence of objects such as birds or cars in an image, translate speech in an audio recording, or detect the presence of a tumour on an X-ray. Labelling data is essential for many machine-learning applications, including machine vision, natural language processing, and speech recognition.

Table 5 shows that the results of data labelling are crucial in building sentiment analysis models that are accurate and relevant to the specific context of cyberbullying. Labelled data allows the model to classify sentiment more accurately. After pre-processing and labelling, clean data was obtained and ready for analysis, consisting of 5000 Instagram user comment datasets. The dataset was classified using the SentiStrength algorithm and lexicon-based methods [14]. SentiStrength generates positive, negative, and neutral values, and the text polarity decision is based on the largest emotion value. In this analysis, the tweet data used is in Indonesian to ensure more accurate classification as it matches the dictionary used. The assessment is done by looking at comments with a positive polarity value.

Table 5. Labelling result of the comment data

Text_Clean	Score	Sentiment
hah sih sok jago	-3	Negative
wadeh nyiapin maaf	-2	Negative
wkwkwk	0	neutral
kasihan anaknya	4	positive

 ISSN:2302-8149
e-ISSN:2540-9719

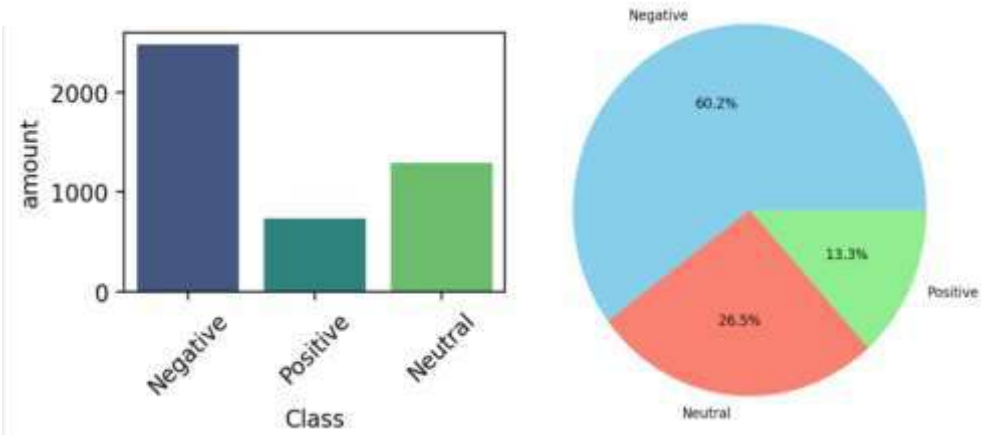

Figure 2. Visualization results evaluation of cyberbullying comments

Figure 2 The sentiment classification results using SentiStrength show that 13.3% of comments are neutral, 13.3% are positive, and 60.2% are negative. From this data, it can be concluded that people's responses to cyberbullying tend to be negative. According to researchers, the high negative value is caused by Instagram users who often misperceive and do not consider carefully before giving an opinion about a post [15].

TF-IDF Weighting

TF-IDF counting is an important technique in natural language processing that combines Term Frequency (TF) and Inverse Document Frequency (IDF) to assess the importance of words in a document. This method is useful for identifying important terms and improving the accuracy of text analysis [16]. Term Frequency (TF) Calculation This process calculates the number of words that appear in the dataset and determines each word/term in the dataset [16]. The number of occurrences of words in one line of the document / the number of words in the document. TF calculations can be found in Table 6.

Table 6. TF weighting table

Steaming	Term Frequency (TF)	
	Word	TF Value
['pikir', 'pikir' 'bayar', 'hutang', 'modal', 'pilpres', 'saja']	*Pikir*	2/6
	bayar	1/6
	hutang	1/6
	modal	1/6
	pilpres	1/6
	saja	1/6
['ubah', 'orang', 'orang']	*ubah*	1/2
	orang	2/2

Inverse Document Frequency (IDF) calculation involves counting the total number of documents in the corpus and determining how many documents contain a particular word. This allows the identification of words that appear infrequently but have high information value in the document collection [17]. The IDF calculation can be found in Table 7

Table 7. IDF weighting table

 ISSN:2302-8149
e-ISSN:2540-9719

Word	N	DF	IDF
Pikir	2	1	0.301029
Bayar	1	1	0
hutang	1	1	0
modal	1	1	0
pilpres	1	1	0
saja	1	1	0
ubah	1	1	0
orang	2	1	0.301029

Inverse Document Frequency (TF-IDF) calculation is a process that produces a final value for each word in the document compared by multiplying the Word Frequency (TF) by the Inverse Document Frequency (IDF). This value indicates the importance of the word in the document compared to the entire corpus, helping to identify important words in text analysis [18]. The calculation of TF - IDF can be seen in Table 8.

Table 8. IDF weighting table

word	N	IDF	TF-IDF
Pikir	2	0. 301029	0.100342
Bayar	0	0	0
hutang	0	0	0
modal	0	0	0
pilpres	0	0	0
saja	0	0	0
ubah	0	0	0
orang	2	0. 301029	0. 100342

Confusion Matrix

After completing all the analysis steps, the final stage involves evaluation using a 3x3 confusion matrix. The model can classify the data into positive, negative, and neutral classifications. Thus, we can evaluate the performance and accuracy of the sentiment analysis model and identify areas where the model needs further improvement [19]. The results of the calculations are displayed in Figure 3, and the accuracy, precision, and recall metrics are listed in Table 9.

 ISSN:2302-8149
e-ISSN:2540-9719

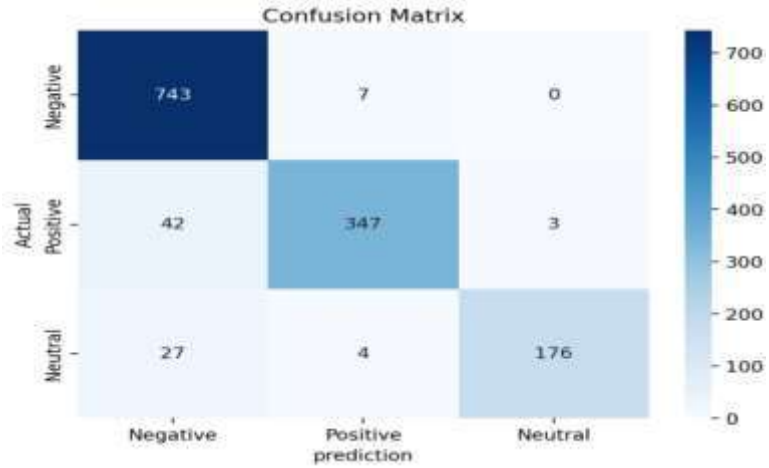

Figure 3. Confusion matrix results in 3 X 3

Table 9. Sentistrenght result

Matrix	Score	Presentasi
Accuracy	0.9300651354130957	93,0%
Precision	0.9651898734177216	96,5%
Recall	0.8840579710144928	88,4%

5. Conclusions

The results show that sentiment analysis is an effective tool for understanding public sentiment, especially on social media platforms such as Instagram, regarding the issue of cyberbullying. The results of the analysis revealed the dominance of negative sentiments with a significant percentage, reaching 62.2%, signalling great dissatisfaction and concern for the issue of cyberbullying among netizens. However, the findings also showed positive (13.3%) and neutral (26.5%) responses, illustrating the complexity of people's perceptions and responses to the phenomenon. In addition, the analysis shows that the Sentistrength method provides satisfactory results in classifying sentiment. The method effectively identified and understood people's sentiments towards cyberbullying with an average accuracy of about 93.0%, an average precision of 96.5%, and an average recall of 88.4%. The findings provide valuable insights for researchers, practitioners, and policymakers to develop more effective cyberbullying prevention and response strategies and increase public awareness and support.

References

[1] T. Sihotang, "Analisis Sentimen Pengguna Twitter Terhadap Kasus Kebocoran Data Masyarakat Indonesia Menggunakan Algoritma Support Vector Machine (Svm)," pp. 1–19, 2023.

[2] H. A. Dewi, S. Suryani, and A. Sriati, "Faktor Faktor yang Memengaruhi *Cyberbullying* Pada Remaja: A Systematic Review," J. Nurs. Care, vol. 3, no. 2, Jun. 2020.

[3] A. Putri and A. Muzakir, "Analisis Sentimen Cyberbullying KPOP di Media Sosial Twitter menggunakan Metode Naive Bayes," J. Syntax Lit., vol. 7, no. 9, 2022, Accessed: May 26, 2023.

[4] M. Rifauddin, "Fenomena Cyberbullying pada Remaja (Studi Analisis Media Sosial Facebook)," Khizanah Al-Hikmah, vol. 4, pp. 35–44, 2016.

[5] N. F. Hasan and Vera Wati, "Deteksi Cyberbullying pada Facebook Menggunakan Algoritma K-Nearest Neighbor", jss, vol. 1, no. 1, pp. 35-44, Sep. 2021.

[6] R. W. Hardian, P. E. Prasetyo, U. Khaira, and T. Suratno, "Analisis Sentiment Kuliah Daring di Media Sosial Twitter selama Pandemi Covid-19 menggunakan Algoritma Sentistrength," MALCOM.,vol. 1, no.5, 2021.

[7] M. F. Rizki, K. Auliasari, and R. P. Prasetya, "Analisis Sentiment Cyberbullying pada Sosial Media Twitter Menggunakan Metode Support Vector Machine," JATI, vol. 5, no. 2, pp. 548–556, 2021.

 ISSN:2302-8149
e-ISSN:2540-9719

[8] P. E. P. U. Septika Sari, Tri Suratno, Ulfa Khaira, "Analisis Sentimen terhadap Komentar Beauty Shaming di Media Sosial Twitter menggunakan Algoritma SentiStrength," IJIRSE., vol. 1, no.3, p. 8, 2021.
[9] S. Gouthami, "Automatic Sentiment Analysis Scalability Prediction for Information Extraction Using SentiStrength Algorithm," Lect. Notes Networks Syst., vol. 612, doi: 10.1007/978-981-19-9228-5_3.
[10] M. Stephenson, "Using SentiWordNet and Sentiment Analysis for Detecting Radical Content on Web Forums," Acad. Manag., vol. 51, no. September, pp. 1–51, 2021.
[11] M. Thelwall, "The Heart and Soul of the Web? Sentiment Strength Detection in the Social Web with SentiStrength," Underst. Complex Syst., vol. 5, pp. 119–134, 2017, doi: 10.1007/978-3-319-43639-5_7.
[12] S. S. Milania, C. Suhery, and T. Rismawan, "Fever Classification Using the Neighbor Weighted K-Nearest Neighbor Method," CESS., vol. 8, no. 2, pp. 250–261, Jul. 2023, doi: 10.24114/CESS.V8I2.43267.
[13] O. Fanny and H. Suroyo, "Analisis Sentimen Pengguna Media Sosial Terhadap Omnibus Law Berdasarkan Hashtag di Twitter Analysis of Social Media Users Sentiments against Omnibus Law Based on Hashtags on Twitter," Sistemasi, vol. 11, no. 1, pp. 197–206, 2022.
[14] E. Miranda, V. Gabriella, S. A. Wahyudi, and J. Chai, "Text Classification untuk Menganalisis Sentimen Pendapat Masyarakat Indonesia terhadap Vaksinasi Covid - 19 Text Classification for Analysing Indonesian People ' s Opinion Sentiment for," J. Sist. Inf., vol. 12, pp. 438–451, 2023.
[15] M. Naufal, B. Balit, F. S. Utomo, I. S. Program, and C. Java, "Sentiment Analysis of pegipegi . com Review on Google Play Store with Naïve Bayes," vol. 13, pp. 1044–1053, 2024.
[16] P. D. Turney, "Mining the Web for Lexical Knowledge to Improve Keyphrase Extraction: Learning from Labelled and Unlabelled Data," 2002.
[17] M. Kitsuregawa., "Modern information retrieval: A brief overview," vol. 24, no. 4, 2003.
[18] L. Havrlant and V. Kreinovich, "A Simple Probabilistic Explanation of Term Frequency-Inverse A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (TF-IDF) Heuristic (and Variations Motivated Document Frequency (TF-IDF) Heuristic (and Variations Motivated by Th," 2014.
[19] N. Hardi, Y. Alkahfi, P. Handayani, W. Gata, and M. R. Firdaus, "Analisis Sentimen Physical Distancing pada Twitter Menggunakan Text Mining dengan Algoritma Naive Bayes Classifier," Sistemasi, vol. 10, no. 1, p. 131, 2021, doi: 10.32520/stmsi.v10i1.1118.