Logistic Regression for Sentiment Analysis of Insecurity Phenomena on Platform X

Emeylia Safitri (1), Wara Alfa Syukrilla (2), Ika Nur Laily Fitriana (3)

[1,3]Statistics Study Program, Faculty of Science and Technology, Universitas Terbuka

[2]Psychology Study Program, The Faculty of Psychology, UIN Syarif Hidayatullah Jakarta

Jl. Cabe Raya, Pondok Cabe, Pamulang, Tangerang Selatan 15437

e-mail: emeylia.safitri@ecampus.ut.ac.id(1), wara.alfa@uinjkt.ac.id(2), ika.nur@ecampus.ut.ac.id(3)

ABSTRAK

Fenomena insecurity sebagai gejala psikologis semakin sering menjadi bahan diskusi di media sosial. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap fenomena insecurity yang diekspresikan melalui unggahan di Platform X. Penggunaan analisis sentimen dalam fenomena insecurity menjadi penting karena fenomena tersebut bersifat subjektif dan seringkali tidak terdeteksi dalam kehidupan nyata. Dalam konteks ini, analisis sentimen menjadi alat yang efektif untuk menggali sentimen pengguna secara sistematis dan objektif. Data dikumpulkan dari tweet berbahasa Indonesia pada Januari 2025 dengan kata kunci terkait seperti "insecure", "minder", dan "overthinking". Setelah melalui proses text preprocessing, data diklasifikasikan menjadi tiga kategori sentimen: positif, netral, dan negatif. Model regresi logistik digunakan sebagai metode klasifikasi dengan validasi silang 10-lipat untuk mengevaluasi kinerja model. Hasil penelitian menunjukkan dominasi sentimen negatif sebesar 73.34%, sedangkan sentimen positif dan netral masing-masing sebesar 20.38% dan 6.28%. Rata-rata akurasi model mencapai 83.13% dengan performa terbaik dalam mendeteksi sentimen negatif. Visualisasi wordcloud memperlihatkan dominasi kata-kata dengan nuansa negatif, seperti "takut", "rendah", dan "sendiri". Temuan ini menyoroti pentingnya pemahaman lebih lanjut terhadap dinamika psikologis masyarakat digital. Penelitian ini juga membuka ruang bagi intervensi berbasis data dalam mendukung literasi kesehatan mental di ranah daring.

Kata kunci: Insecurity, Regresi Logistik, Analisis Sentimen, Platform X, Text Mining

ABSTRACT

The phenomenon of insecurity as a psychological symptom is increasingly becoming a topic of discussion on social media. This study aims to analyze public sentiment toward the phenomenon of insecurity as expressed through posts on Platform X. The use of sentiment analysis in the context of insecurity is crucial because the phenomenon is subjective and often undetectable in real life. In this context, sentiment analysis is an effective tool for systematically and objectively exploring user sentiment. Data was collected from Indonesian-language tweets in January 2025 using related keywords such as "insecure","minder", and "overthinking." After undergoing text preprocessing, the data was classified into three sentiment categories: positive, neutral, and negative. Logistic regression was employed as the classification method, with 10-fold cross-validation used to evaluate model performance. The study's results show a dominance of negative sentiment at 73.34%, with positive and neutral sentiments accounting for 20.38% and 6.28%, respectively. The model's average accuracy reached 83.13%, with the best performance in detecting negative sentiment. Wordcloud visualizations revealed a dominance of negatively nuanced words such as "takut","rendah" and "sendiri." These findings underscore the importance of deeper understanding of the psychological dynamics of the digital public. This study also paves the way for data-driven interventions to support mental health literacy in online spaces.

Keywords: *Insecurity; Logistic Regression; Sentiment Analysis; Platform X; Text Mining*

INTRODUCTION

The development of information and communication technology has brought major transformations in the way individuals interact and express their opinions. Social media has become the primary channel for voicing perspectives on various social, political, and psychological issues. Among the available platforms, Platform X (previously known as Twitter) stands out as one of the most dynamic media with high levels of public participation [1]. In January 2024, Indonesia had approximately 27.5 million X users, making it the fourth largest user base in the world after the United States, Japan, and India. Its widespread use in Indonesia makes the platform a representative reflection of digital society expression.

Platform X has become a representative space for digital expression in Indonesia, whether in political [2], educational [3], or socio-economic [4] issues. One psychological phenomenon frequently appearing in online discourse is insecurity — an emotional condition involving feelings of insecurity, low self-esteem, or lack of confidence. This condition can cause significant stress and have a negative impact on an individual's mental well-being and performance in the workplace [5]. This phenomenon not only reflects individual dynamics but also forms part of the collective narrative shaped in online social interactions. Intensive use of social media among adolescents, especially females, has been identified as a major trigger of feelings of insecurity, resulting from social comparisons to standards of perfection displayed on social media. In one study, 83.5% of young women aged 18–21 experienced insecurity due to social media, particularly Instagram, and they demonstrated strong dependency on such platforms in their daily lives [1].

In the fields of statistics and data science, social media serves as an unstructured data source that poses analytical challenges. Various approaches such as text mining, sentiment analysis, and statistical modeling have been developed to extract meaningful insights from this social data. In this context, text mining becomes the key to extracting useful knowledge from this unstructured text data [6]. Logistic regression model based on Twitter data could achieve up to 89.83% accuracy in classifying public sentiment regarding face-to-face learning [3]. Furthermore, logistic regression and found that location and business activities significantly affect MSME income in Surabaya [4]. The logistic regression method was chosen because it is effective in classifying sentiments into discrete categories such as positive, negative, and neutral, and easy to interpret. This method aligns well with data represented numerically, such as TF-IDF, and is suitable for relatively simple text datasets. Besides being computationally efficient, logistic regression also allows for analysis of the influence of each word feature on the classification results. This supports the research objective of understanding sentiment patterns toward insecurity phenomena in a measurable and in-depth manner.

Logistic regression has also been applied to socio-economic studies [7], found that individuals who were not household heads (OR = 8.14), unmarried (OR = 4.48), and had lower education (OR = 4.47) were at higher risk of unemployment in West Java. Additionally, use of Platform X significantly increased online (R^2 = 12.7%) and offline (R^2 = 6.7%) political participation, with a strong correlation between the two (r = 0.698) [2]. In previous research, sentiment analysis was used to examine public sentiment towards data security issues (close to psychological insecurity towards information) using machine learning with data sources from platform X [8]. Similar research regarding sentiment analysis towards insecurity phenomena has not been widely found.

However, specific studies exploring the insecurity phenomenon on Indonesian social media using text-based sentiment analysis approaches remain limited. Text mining is not only a tool for understanding what people are talking about, but also a key to designing more effective strategies to respond to the dynamics that occur in cyberspace [8]. This study aims to apply sentiment analysis to posts about insecurity on Platform X using logistic regression. Additionally, this research evaluates classification model performance via cross-validation and confusion matrix analysis, while illustrating Indonesian public sentiment patterns related to insecurity. It is expected that the results will contribute scientifically to digital literacy development, increased mental health awareness, and data-based social policy formulation in Indonesia.

METHOD

This study adopts a quantitative content analysis approach based on text mining techniques to explore and analyze public sentiment regarding the phenomenon of insecurity — feelings of uncertainty — within various social, political, and economic contexts. Text mining is a data science method used to extract valuable information from collections of unstructured text data [10]. Text mining is not only a tool for understanding what people are talking about, but also a key to designing more effective strategies to respond to the dynamics that occur in social media [9] This technique enables the identification of patterns, trends, and hidden opinions within large volumes of text data, making it highly relevant for studying public perception dynamics on social media. This study used logistic regression to model the relationship between a categorical dependent variable (multinominal) and one or more independent variables, which can be continuous or categorical. Unlike linear regression, which predicts numerical values, logistic regression predicts the probability of an event or a specific category using the following formula:

$$(Y = 1 \mid X) = \frac{1}{(1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)})} \quad (1)$$

Where, P(Y=1|X) is probability of the event occurring (Y = 1) given the predictor variables X, β_0 is intercept (log-odds when all predictors are 0), β_i is regression coefficients representing the influence of each predictor, and X_i is predictor variables.

The stages of data analysis in this study are as follows:

1. Data Collection

   The data collection process began with the retrieval of tweets related to the topic of insecurity using keywords such as “insecure”, “minder” “insecurity” and “overthinking.” The collection focused on Indonesian-language tweets during January 2025 and filtered for geolocation within Indonesia to ensure cultural relevance. A total of 5,226 tweets were collected for analysis.
2. Text Processing

   Raw data was cleaned from irrelevant elements such as symbols, usernames, URLs, punctuation, and numbers. This included converting text to lowercase (case folding), tokenization into individual words, and stopword removal to retain only meaningful terms. These steps aimed to simplify the text and improve the quality of feature representation.
3. Sentiment Labeling

   After cleaning, the tweets were manually labeled based on their semantic content. Each tweet was classified into one of three sentiment categories—positive, negative, or neutral—

considering the explicit and implicit context. Previous research also carried out manual labeling by researchers to obtain accurate results [11].

4. Wordcloud Visualization
   To obtain an initial view of frequently occurring words, a wordcloud visualization was generated. This allowed identification of dominant keywords related to the insecurity phenomenon and supported the interpretation of sentiment.
5. Text Transformation into Numerical Representation
   Normalized text was then transformed into numerical form using the Term Frequency–Inverse Document Frequency (TF-IDF) method [12]. This representation enables statistical algorithms to recognize the importance of each word across the corpus. The dataset was split into training and testing sets for model training and performance evaluation.
6. Modeling and Performance Evaluation
   Logistic regression was used as the classification algorithm to map tweets into sentiment categories based on the generated feature representations. Model evaluation used k-fold cross-validation (k=10) to avoid overfitting and obtain stable performance estimates. One of the most widely used methods to assess the performance of a classification model is the confusion matrix or classification table [9]. Evaluation metrics included accuracy, precision, recall, and F1-score, calculated from the confusion matrix of classification results.

RESULT AND DISCUSSION

The dataset analyzed consists of 5,226 tweets collected from Platform X during January 2025. Tweets were classified into three sentiment categories—positive, neutral, and negative—based on semantic analysis and manual labeling. The classification results showed that 1,126 tweets (20.38%) were positive, 347 tweets (6.28%) were neutral, and 4,053 tweets (73.34%) were negative. This distribution is visualized in Figure 1.

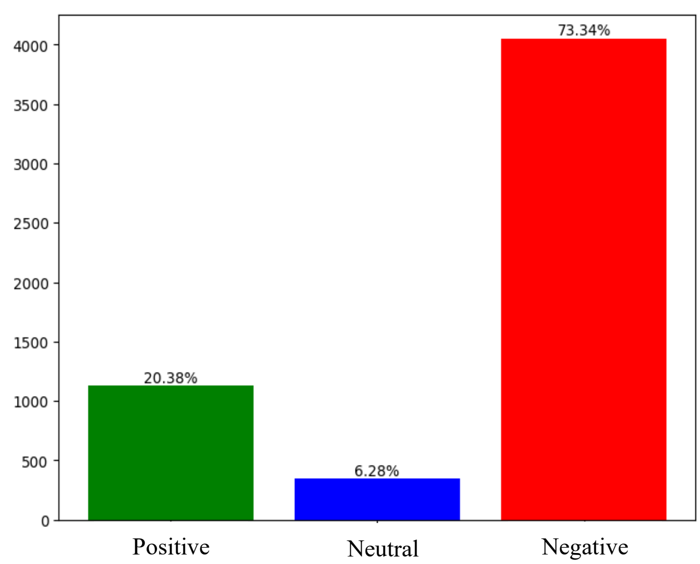

Figure 1. Sentiment visualization

The wordcloud in Figure 2 shows that words such as "merasa" and "diri" appear most prominently, indicating high frequency in the tweet corpus. This dominance suggests that feelings of insecurity are closely linked to individual perceptions of the self. The prominence of the word "percaya" also reflects a need for self-validation and efforts to rebuild confidence. The terms "overthinking" and "minder" highlight excessive rumination and low self-esteem as central themes in discussions of insecurity. Words such as "tidak" "takut" "sendiri" and "rendah" further reinforce the negative emotional spectrum coloring this discourse. On the other hand, words such as "benar" "percaya" "semangat" and "mampu" indicate the presence of affirming narratives and optimistic encouragement in overcoming insecurity. This reflects a dialectic between vulnerability and resilience within public discourse on social media. Overall, the wordcloud visualization provides strong insights into the psychological structure of users expressing insecurity, revealing a complex narrative of uncertainty, self-discovery, and intense emotional dynamics within digital society.

Figure 2. Word Cloud of Dominant Words in Tweets Based on Sentiment Category

Before classification modeling is performed, text data needs to be cleaned and prepared so that it can be represented in a numeric form that matches the input format of the machine learning algorithm. This preprocessing process aims to remove irrelevant elements and standardize the text to be optimal for further analysis. The initial step includes removing usernames (the "@" character), URLs (links with the prefix "http" or "www"), and hashtags (the "#" character and the word after it). These elements generally have no informative value in the context of sentiment analysis, and have the potential to be noise in modeling. In addition, emoticons, non-alphabetic symbols, punctuation, and numbers are also removed to maintain text consistency.

After preprocessing, the text is converted into numeric form using CountVectorizer, which produces a document-term matrix (DTM). This DTM is the basis for classification modeling using logistic regression, where each tweet is represented as a numeric vector based on the frequency of word occurrence. To evaluate model performance objectively and avoid overfitting, the k-fold cross-validation technique is used with a value of $k = 10$. This method divides the data into 10 random subsets, and in each iteration, one subset is used as test data while the remaining nine are used as training data, the evaluation results of which can be seen in Figure 3. Based on Figure 3, the accuracy scores of each fold are in a relatively narrow and consistent range, without extreme fluctuations. The average cross-validation accuracy was recorded at 83.13%, as indicated by the red horizontal line on the graph. This value reflects that the logistic regression model has a good and stable classification performance in predicting sentiment based on text features.

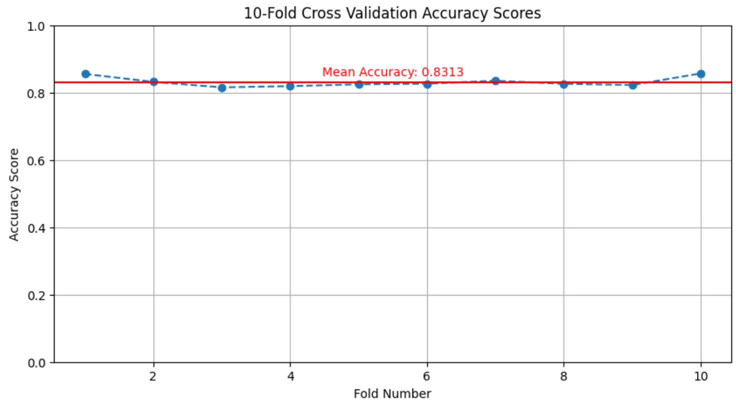

Figure 3. Accuracy Score on 10-Fold Cross Validation Using Logistic Regression Model

Based on Figure 4, the results of the model performance evaluation using the confusion matrix show that the logistic regression model has very good classification performance in detecting negative sentiment. From a total of 816 actual data with negative labels, the model successfully classified 800 data correctly, so that the accuracy for the negative class reached 98%. This shows that the model has a high ability to detect negative opinions on social media. However, the model's performance for the neutral and positive classes was much lower. Only 39% of neutral data (28 out of 71) and 45% of positive data (99 out of 219) were successfully classified correctly. Most of the misclassifications occurred in neutral and positive data that were misclassified as negative, as many as 38 and 115 cases, respectively.

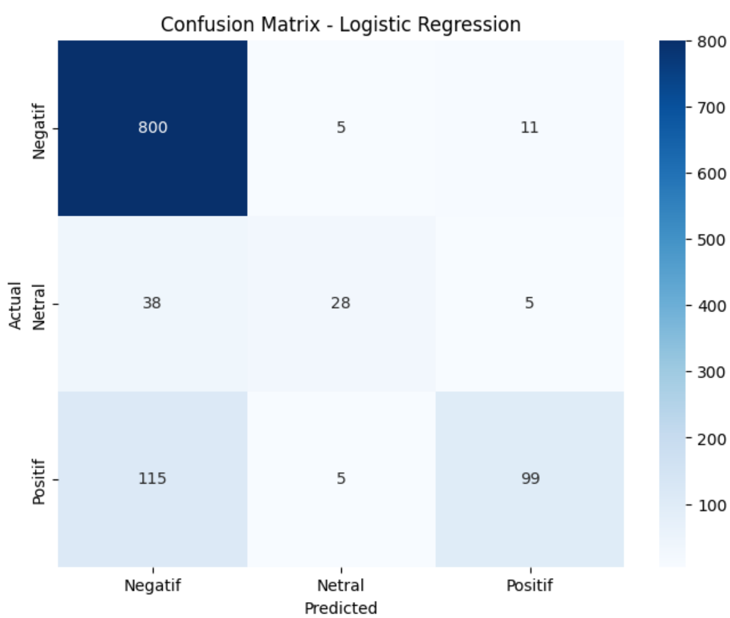

Figure 4. Confusion Matrix

Table 1 Classification Report

Class	Precision	Recall	F1-Score	Support
Negative	0.84	0.98	0.9	816
Neutral	0.74	0.39	0.51	71
Positive	0.86	0.45	0.59	219
Accuracy			**0.84**	**1106**

Confussion matrix in Figure 4 and classsification report in Tabel 1 are the results of modeling 80% of the data, then 20% becomes the test data. Meanwhile, Based on Table 1, it can be seen that the testing of the logistic regression model on the test data produces a confusion matrix that describes the distribution of predictions for three sentiment classes, namely Negative, Neutral, and Positive. The evaluation results show that the model has the best performance in detecting the Negative class. Of the total 816 actual data in the Negative class, the model successfully classified 800 data correctly, resulting in a recall value of 98.04%. This shows that the model has a very high sensitivity to Negative sentiment. In contrast, the model's performance on the Neutral class is moderate. Out of 71 actual Neutral data, only 28 data were predicted correctly, while 38 other data were classified as Negative, and 5 other data were classified as Positive. For the Positive class, the model shows a fairly good level of accuracy, although it is not yet optimal.

CONCLUSION

This study successfully revealed public perceptions of the insecurity phenomenon on Platform X using logistic regression-based sentiment analysis. The main finding shows a strong dominance of negative sentiment (73.34%), significantly higher than positive (20.38%) and neutral (6.28%) sentiments. This indicates that insecurity is generally perceived as a negative psychological experience within Indonesia's digital society. The wordcloud visualization reinforced this by highlighting dominant words such as “afraid,” “inferior,” and “overthinking,” which reflect users’ emotional vulnerability.

From a modeling perspective, logistic regression demonstrated good classification performance, with an average accuracy of 83.13%. However, confusion matrix evaluation revealed limitations in detecting minority classes. Low recall values for neutral (0.39) and positive (0.45) classes indicate the presence of majority class bias, a common challenge in imbalanced datasets. Misclassification of 115 positive and 38 neutral tweets as negative also reduced the model’s sensitivity, especially in early detection applications that require high precision. Academically, this study contributes to the literature on text mining applications in mental health issues in Indonesia and highlights technical challenges in sentiment classification using social media data.

Future studies are recommended to address class imbalance using techniques such as oversampling, under sampling, or class weighting to improve performance in minority classes [14]. More advanced text representations like word embeddings (e.g., Word2Vec or Fast Text), or deep learning approaches based on transformer models (such as BERT), are also suggested to enhance contextual understanding. Research can also be expanded by increasing the dataset size, extending the data collection period, or focusing on specific regions with higher insecurity prevalence. Practically, these findings may benefit mental health practitioners, policymakers, and social organizations in designing more targeted and data-driven intervention strategies.

ACKNOWLEDGMENTS

The author would like to thank the Statistics Study Program, Faculty of Science and Technology, Universitas Terbuka and Psychology Study Program, The Faculty of Psychology, UIN Syarif Hidayatullah Jakarta, as well as all parties involved in supporting this research and publication.

REFERENCE

[1] A. Valentina, G. L. Putri, V. Valiani, and O. H. Putri, "Komunikasi Visual untuk Edukasi Insecurity pada Remaja Perempuan yang Diakibatkan oleh Penggunaan Media Sosial,", *bahasarupa*, vol. 5, no. 2, pp. 237–245, 2022.

[2] D. D. Anwar, K. A. Wibowo and H. S. M. Rohayati, "Pengaruh Penggunaan Media Sosial X (Twitter) Terhadap Partisipasi Politik pada Pemilu Presiden Indonesia Tahun 2024," *Jurnal Riset Komunikasi*, vol. 15, no. 2, pp. 314–328, 2024.

[3] S. A. R. Manaf, A. Alamudi, and A. Fitrianto, "Sentiment Analysis of Twitter Users' Opinion Towards Face-to-Face Learning: Analisis Sentimen Tanggapan Masyarakat Pengguna Twitter terhadap Pembelajaran Tatap Muka,", *Indonesian Journal of Statistics and Its Applications*, vol. 7, no. 1, pp. 15–31, 2023.

[4] A. N. Alifah, A. I. Edina, and M. Almuhayar, "Ordinal Logistic Regression Model of Micro, Small, and Medium-Sized Enterprises Income: A Case Study of Micro, Small and Medium-Sized Enterprises in Surabaya,", *Indonesian Journal of Statistics and Its Applications*, vol. 8, no. 2, pp. 143–154, 2024.

[5] C. Lee, G. H. Huang, S. J. Ashford, Jayanti, S. P. Palupi, and K. A. Notodiputro, "Job Insecurity and the Changing Workplace: Recent Developments and the Future Trends in Job Insecurity Research,", *Annual Review of Organizational Psychology and Organizational Behaviour*, vol. 5, no 1, pp. 335–359, 2018.

[6] R Nikhil, N. Tikoo, S. Kurle, H. S. Pisupati, and G. R. Prasa, "Sentiment Analysis of Twitter Users' Opinion Towards Face-to-Face Learning,", *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 2, no. 4, pp. 1292-1296, 2015.

[7] D. Jayanti, S. P. Palupi, and K. A. Notodiputro, "A Conditional Logistic Regression Model for Analyzing Unemployment Rates in West Java: Model Regresi Logistik Bersyarat untuk Analisis Tingkat Pengangguran di Provinsi Jawa Barat,", *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 195–204, 2021.

[8] A. Z. Taufan and W. Wibowo, "Analisis Sentimen Terkait Persepsi Keamanan Data Informasi Dan Privasi di Indonesia Menggunakan Pendekatan Machine Learning,", *JINTEKS*, vol. 6, no. 3, pp. 728-736, 2024.

[9] Y. Zhang, Z. Xu, and J. Lin, "Text Mining for Social Media Analysis: Trends and Techniques,", *Journal of Data Science*, vol. 20, no. 1, pp. 65–80, 2022.

[10] S. M. Mohammad. "Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text,". *In H. L. Meiselman (Ed.), Emotion measurement (2nd ed.) Elsevier*, pp. 323–379), 2021.

[11] A. F. Ningrum and I. F. Amri, "Sentiment Analysis Of Public Opinion On Handling Stunting In Indonesia Using Random Forest,", *J Statistika*, vol. 17, no. 1, pp. 645–654, 2024.

[12] D. K. A. Astutik, A. Indrasetianingsih, F. Fitriani, "Penerapan Text Mining pada Analisis

Sentimen Pengguna Twitter Layanan Transportasi Online Menggunakan Metode Density Based Spatial Clustering of Applications With Noise (DBSCAN) dan K-Means,", *J Statistika*, vol. 15, no. 1, pp. 184-194, 2022.

[13] R. Singh, and A. Gupta, "Precision and Recall in Machine Learning Algorithms,", *Journal of Machine Learning Applications*, vol. 11, no. 3, pp. 213–221, 2020.

[14] N. Thamrin, A. F. Baktiar, F. A. Addawiyah, M. Husna, and T. Irwati, "Determinants of Antenatal Care Visits in Indonesia with Synthetic Minority Over-Sampling Techniques for Imbalance Data: Determinan Kunjungan Antenatal Care di Indonesia dengan Teknik Synthetic Minority Over-Sampling untuk Imbalanced Data,", *Indonesian Journal of Statistics and Its Applications*, vol. 7, no. 2, pp. 86–104, 2023.