RESEARCH ARTICLE

BORNEO SCIENCE
THE JOURNAL OF SCIENCE & TECHNOLOGY

Journal homepage: https://jurcon.ums.edu.my/ojums/index.php/borneo-science

SENTIMENT ANALYSIS OF PUBLIC HEALTH SOCIAL MEDIA COMMENT USING EXPERT ANNOTATION

Daimler B. Alebaba, Suaini Sura*, Nooralisa M. Tuah, and Nona M. Mohd Nistah

Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jln UMS Kota Kinabalu 88400, Sabah, Malaysia.

***Correspondence:**
su_sura@ums.edu.my

Received: 29 April 2025
Revised: 6 January 2026
Accepted: 6 January 2026
Published online: 21 January 2026

Doi:
10.51200/bsj.v46i2.6321

Keywords:
Expert annotation; Public health; Sentiment analysis; Social media; User engagement.

ABSTRACT. *Sentiment analysis has become a critical tool for organizations and researchers to understand user sentiment. However, it faces challenges such as managing noisy data, interpreting sarcasm or irony, and adapting to the evolving nature of language, especially in public health, where users often express opinions about their health conditions, healthcare experiences, and complex medical terminology on various topics. Addressing these challenges is crucial to maintaining the integrity of sentiment analysis results. Hence, this study analyzes public health social media user comments using a structured sentiment analysis framework. The framework includes dataset collection and annotation, text preprocessing, feature vectorization, and text classification. Experimental results show that the proposed model achieved an accuracy of 98% on the annotated dataset, indicating strong predictive performance within the scope of this study. The findings suggest that the framework is effective in capturing sentiment patterns in public health social media data and provides a foundation for further evaluation using comparative and benchmark-based analyses.*

INTRODUCTION

Sentiment analysis is a subfield of Natural Language Processing (NLP) and has become a vital method for organizations and researchers seeking to understand user sentiment (Aftab *et al*., 2023; Hartmann *et al*., 2023). The origins of sentiment analysis trace back to early research in 2001, when Das and Chen attempted to use evaluative texts to predict public sentiment, laying the groundwork for what would become a rapidly expanding field of study (Ogbuokiri *et al*., 2024; Venkit *et al*., 2023). Since the foundational work of Das and Chen, sentiment analysis has undergone significant advancements. Researchers began developing methods to classify the polarity of text, which involves determining whether a given piece of text expresses a positive or negative sentiment (Tan *et al*., 2023). This evolution was driven by the increasing availability of digital text data, particularly from public health social media platforms where users frequently express their opinions concerning their health condition, healthcare, and complex medical-related terminology on a variety of topics.

However, sentiment analysis faces several challenges, including managing noisy data, interpreting sarcasm or irony, and adapting to the fluidity of language (Lakshmi *et al*., 2024, 2024). These challenges could lead to misleading insights and incorrect sentiment interpretation, resulting in faulty decision-making based on inaccurate data. In the context of public health, for example, inaccurate sentiment analysis might cause misunderstandings of user concerns or needs, leading to ineffective responses or interventions (Mohammad Amini *et al*., 2023). Furthermore, poor accuracy can reduce the reliability of predictive models, limiting their practical applications and diminishing trustworthiness in real-world scenarios (Adli *et al*., 2024; Lakshmi *et al*., 2024). Therefore, addressing challenges such as noisy data and sarcasm is essential to maintain the integrity of sentiment analysis results.

Hence, this study analyzes public health social media user comments using a structured sentiment analysis framework. The framework includes the steps of dataset collection and annotation, text preprocessing, feature vectorization, and text classification (Adli *et al*., 2024; Aftab *et al*., 2023; Md Suhaimin *et al*., 2019). Each step is designed to refine and ensure the relevance of the data, enhancing the accuracy of text classification (Venkit *et al*., 2023). The final step involves evaluating the accuracy of the classification results.

MATERIALS AND METHODS

The method of this study follows a structured sentiment analysis framework, which includes annotation by three public health experts, text preprocessing using NLP techniques, feature vectorization with Term Frequency-Inverse Document Frequency (TF-IDF), and classification using the Support Vector Machine (SVM) algorithm. Classification accuracy is evaluated using metrics such as the confusion matrix and assessed through the Area Under the Curve (AUC). The framework of this study is illustrated in Figure 1.

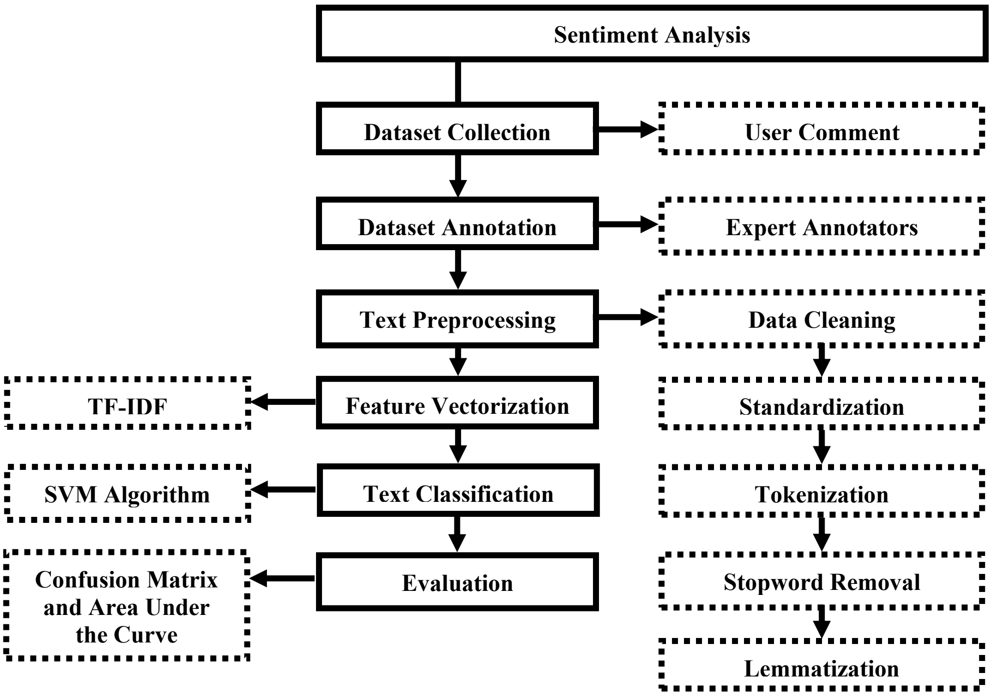

Figure 1. Structure sentiment analysis framework.

Dataset Collection

The dataset was sourced from the public health Facebook page of Hospital UMS, comprising user comments in textual data. Permission to collect this dataset was formally granted by Hospital UMS, ensuring adherence to ethical standards and data protection regulations (Mohammad Amini *et al*., 2023). The dataset spans from June 1, 2020, to September 30, 2022, and was stored in a Microsoft Excel worksheet. Given that, mixed languages, including Malay and English, appeared in the dataset. To ensure consistency, the Google Translate Sheets function was used to translate all non-English comments into English.

Dataset Annotation

Three public health experts were assigned to manually label the sentiment of each comment as either positive or negative. The first annotator is a medical officer, while the second and third annotators are dental surgery assistants. These healthcare professionals were selected for their expertise in medical terminology, health-related issues, and public health concerns. The dataset contains 368 rows and seven columns: id, comment, comment translated, annotator_1, annotator_2, annotator_3, and result. The id column represents the content ID, comment holds the original user comment, and comment_translated contains the translated version in English. The annotator_1 column reflects the sentiment labels assigned by the medical officer, while annotator_2 and annotator_3 correspond to the labels given by the two dental surgery assistants. The result column indicates the inter-annotator agreement on the sentiment polarity.

Text Preprocessing

While the collected and annotated dataset is labeled, it may still contain unnecessary elements that do not contribute to sentiment analysis tasks. Therefore, it is crucial to remove these irrelevant elements. For that, text preprocessing using NLP techniques is conducted to prepare the dataset. This preprocessing involves several stages, including data cleaning, standardization, tokenization, stopword removal, and lemmatization (Ogbuokiri *et al*., 2024). Each stage simplifies and standardizes the text, ensuring a consistent and well-structured dataset that facilitates accurate feature vectorization and text classification (Bordoloi & Biswas, 2023).

Data cleaning removes non-useful elements such as names, emojis, links, and administrative comments from the dataset. This manual process ensures that only relevant user comments are included, refining the dataset and improving the quality of the sentiment analysis. Standardization involves converting all text to lowercase, ensuring uniformity and preventing case-related discrepancies, which facilitates more accurate text classification. This stage is implemented using Python's Pandas library. Tokenization breaks the text into individual words or tokens, transforming sentences into analyzable units, which simplifies the analysis of word frequencies. Stopword removal eliminates common, non-informative words like "and," "the," or "is," as well as numbers and unnecessary punctuation. This reduces the dataset's dimensionality and focuses on words that contribute meaningful information. Lemmatization reduces words to their base or canonical forms using language-specific rules, enhancing accuracy by unifying different word forms into a single representation.

Feature Vectorization

Feature vectorization converts text data into numerical features necessary for predictions (Lakshmi *et al*., 2024; Mohammad Amini *et al*., 2023). In this study, TF-IDF vectorization was used. The dataset, comprising 368 text documents, was split into 70% for training and 30% for testing using the `train_test_split` function from Python's `sklearn.model_selection` library. This ensures random partitioning while maintaining the target variable's distribution. TF-IDF vectorization was then applied to the training set, calculating scores that reflect each term's importance within a document relative to the entire corpus.

Text Classification

Text classification was performed using the SVM algorithm from Python's sklearn.svm library. A polynomial kernel (kernel='poly') was chosen to handle non-linear separations in the data, allowing the SVM to find a flexible decision boundary. The classifier was trained on the TF-IDF vectorized training set using the fit method, which optimizes the hyperplane to maximize the margin between classes and improve classification accuracy.

Evaluation

The evaluation was conducted using the `sklearn.metrics` library in Python. A classification report was generated with the `classification_report` function, detailing performance metrics such as classification accuracy, precision, recall, F1-score, and support for each class. The confusion matrix was used to evaluate and summarize prediction results, showing counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Additionally, the classification accuracy was assessed using the AUC metric, with higher scores indicating better predictive performance. Table 1 outlines the interpretation of AUC scores.

Table 1. Area under the curve.

Score	Predictive Performance
0.90-1.00	Excellent Classification
0.80-0.90	Good Classification
0.70-0.80	Fair Classification
0.60-0.70	Poor Classification
0.50-0.60	Failure

RESULTS AND DISCUSSIONS

The evaluation begins with analyzing both the 70% training set and the 30% testing set using the confusion matrix. The confusion matrix categorizes the predicted instances into four outcomes, namely TP, TN, FP, and FN. The results are summarized in Table 2, highlighting the model's performance in terms of correct and incorrect classifications.

Table 2. Confusion matrix result.

Predicted Instance	Actual Positive	Actual Negative	Metrics
Positive	73 (TP)	1 (FP)	0.99 (precision)
Negative	7 (FN)	51 (TN)	0.00
Metrics	0.91 (recall)	0.98 (specificity)	0.94 (accuracy) and 0.95 (F1-score)

Table 2 reveals that the model has a high precision of 0.99, indicating that 99% of instances predicted as positive were indeed positive. This demonstrates the model's strong predictive performance. The recall of 0.91 suggests that the model correctly identified all actual positive instances without any missing (FN = 7), reflecting its high sensitivity to positive cases. The confusion matrix is visually represented as a heatmap in Figure 2.

The model's specificity is 0.98 due to the absence of true negative instances (TN = 51), indicating that the model did not identify any actual negative cases. This suggests the model is heavily focused on detecting positive instances, potentially at the expense of identifying negatives. Despite this, the model's overall accuracy is 0.94, meaning 94% of all predictions, whether positive or negative, were correct. The F1-score of 0.95 combines precision and recall into a single metric, reflecting a high level of accuracy in positive predictions.

The model's classification accuracy of 0.94 includes a 6% error margin, represented by FP =1. This error margin is within acceptable limits for most practical applications. Achieving absolute perfection is rare in machine learning, and a small number of errors are generally acceptable, especially when the model shows strong precision and recall (Ogbuokiri *et al*., 2024). The classification report, detailed in Table 3, provides a comprehensive breakdown of performance metrics.

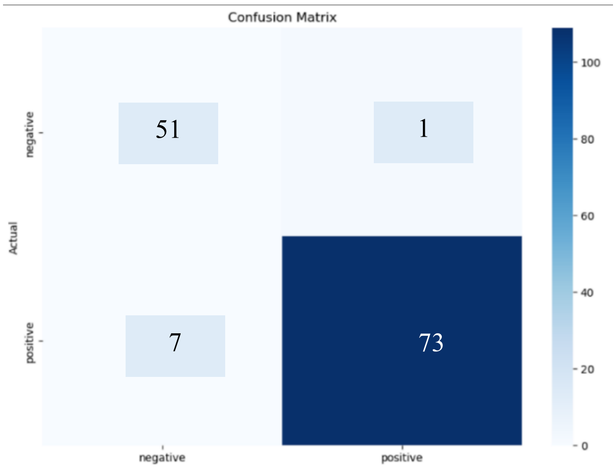

Figure 2. Confusion matrix results in a heatmap plot.

Table 3. Sentiment analysis classification report.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.879310	0.980769	0.927273	52.000000
Positive	0.986486	0.912500	0.948052	80.00000
Accuracy	0.939394	0.939394	0.939394	0.939394
Macro Average	0.932898	0.946635	0.937662	132.000000
Weighted Average	0.944266	0.939394	0.939866	132.000000

According to the AUC evaluation criteria outlined in Table 1, the model's classification accuracy of 0.94 falls within the 'Excellent Classification' range, which is defined as an AUC score between 0.90 and 1.00. This range is widely recognized as an industry standard for high-performing models, indicating that the model is not only accurate but operates with a level of precision and recall that exceeds the benchmarks typically required in practical applications (Lakshmi *et al*., 2024). Models within this range are capable of distinguishing between positive and negative instances with a high degree of certainty, minimizing both false positives and false negatives (Ogbuokiri *et al*., 2024).

.

CONCLUSION

This study successfully analyzed public health social media user comments using a structured sentiment analysis framework. The framework involved annotation by three public health experts, text preprocessing with NLP techniques, feature vectorization using TF-IDF, and classification with the SVM algorithm. The classification achieved 98% accuracy, demonstrating a strong predictive performance. According to the AUC evaluation criteria, the accuracy of 0.94 falls within the 'Excellent Classification' range, where this range is widely recognized as an industry standard for high-performing models, indicating that the model is not only accurate but operates with a level of precision and recall that exceeds the benchmarks typically required in practical applications. The key contribution of this study is that the framework can be used to conduct sentiment analysis in challenging areas such as handling noisy data, detecting sarcasm or irony, and adapting to the fluid nature of language. Future work will focus on comparing the SVM algorithm with Naïve Bayes, Decision Tree, and Neural Networks to explore further improvements in accuracy performance.

ACKNOWLEDGEMENT

The authors wish to express their sincere gratitude to Universiti Malaysia Sabah for the support in facilitating this study (SBK0461-2021, Skim Penyelidikan Bidang Keutamaan). Special thanks are also extended to Hospital Universiti Malaysia Sabah for their contribution to the data collection process, which was crucial for the completion of this study.

REFERENCES

Adli, N.B.Z., Ahmad, M., Ghani, N. A., Ravana, S.D. & Norman, A.A. 2024. An ensemble classification of mental health in Malaysia related to the COVID-19 pandemic using social media sentiment analysis. *KSII Transactions on Internet and Information Systems*, 18(2): 370–396. https://doi.org/10.3837/TIIS.2024.02.006

Aftab, F., Bazai, S.U., Marjan, S., Baloch, L., Aslam, S., Amphawan, A. & Neo, T.-K. 2023. A comprehensive survey on sentiment analysis techniques. *International Journal of Technology*, 14(6): 1288–1298. https://doi.org/10.14716/ijtech.v14i6.6632

Bordoloi, M. & Biswas, S.K. 2023. Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, 56(11): 12505–12560. https://doi.org/10.1007/s10462-023-10442-2

Hartmann, J., Heitmann, M., Siebert, C. & Schamp, C. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1): 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Lakshmi, C.S., Saxena, S. & Kumar, B.S. 2024. Sentiment analysis and classification of COVID-19 tweets using machine learning classifier. *Journal of Autonomous Intelligence*, 7(2): 1–13. https://doi.org/10.32629/jai.v7i2.801

Md Suhaimin, M.S., Ahmad Hijazi, M.H., Alfred, R. & Coenen, F. 2019. Modified framework for sarcasm detection and classification in sentiment analysis. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(3): 1175. https://doi.org/10.11591/ijeecs.v13.i3.pp1175-1183

Mohammad Amini, M., Jesus, M., Fanaei Sheikholeslami, D., Alves, P., Hassanzadeh Benam, A. & Hariri, F. 2023. Artificial intelligence ethics and challenges in healthcare applications: a comprehensive review in the context of the European GDPR Mandate. *Machine Learning and Knowledge Extraction*, 5(3): 1023–1035. https://doi.org/10.3390/make5030053

Ogbuokiri, B., Ahmadi, A., Nia, Z.M., Mellado, B., Wu, J., Orbinski, J., Asgary, A. & Kong, J. 2024. Vaccine hesitancy hotspots in Africa: an insight from geotagged twitter posts. *IEEE Transactions on Computational Social Systems*, 11(1): 1325–1338. https://doi.org/10.1109/TCSS.2023.3236368

Tan, Y.Y., Chow, C.-O., Kanesan, J., Chuah, J.H. & Lim, Y.L. 2023. Sentiment analysis and sarcasm detection using deep multi-task learning. *Wireless Personal Communications*, 129(3): 2213–2237. https://doi.org/10.1007/s11277-023-10235-4

Venkit, P.N., Srinath, M., Gautam, S., Venkatraman, S., Gupta, V., Passonneau, R.J. & Wilson, S. 2023. *The sentiment problem: a critical survey towards deconstructing sentiment analysis*. 13743–13763.