JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY
ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, **2026**, 26, 030-036

PHÂN TÍCH CẢM XÚC CÁC ĐÁNH GIÁ ỨNG DỤNG BẢO HIỂM Y TẾ SỬ DỤNG NAÏVE BAYES, SUPPORT VECTOR MACHINE VÀ SMOTE

Khâu Văn Bích[1], Võ Minh Tiến[2], Trần Hữu Duật[2*]

[1]*Trường Đại học Trần Đại Nghĩa, Số 189 Nguyễn Oanh, Phường 10, Gò Vấp, Hồ Chí Minh, Việt Nam*
[2]*Trường Đại học Thủ Dầu Một, Số 6 Trần Văn Ơn, Phường Phú Lợi, Hồ Chí Minh, Việt Nam*
* Tác giả liên hệ: *duatth@tdmu.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận: 06/03/2026
Ngày hoàn thiện: 27/03/2026
Ngày chấp nhận: 07/04/2026
Ngày đăng: 15/04/2026

TỪ KHÓA

Sentiment analysis;
Naïve Bayes;
Support Vector Machine;
SMOTE ;
Health Insurance Application.

TÓM TẮT

Nghiên cứu này khảo sát các yếu tố ảnh hưởng đến trải nghiệm người dùng đối với ứng dụng bảo hiểm y tế trên thiết bị di động thông qua phân tích đánh giá của người dùng bằng phương pháp phân loại cảm xúc. Hai mô hình phân loại, bao gồm Naïve Bayes (NB) và Support Vector Machine (SVM), được sử dụng để so sánh hiệu quả, trong đó NB cho thấy kết quả vượt trội hơn về độ chính xác (precision) và chỉ số F1-score. Kết quả cũng cho thấy bước tiền xử lý dữ liệu có vai trò quan trọng đối với hiệu suất mô hình. Cụ thể, việc áp dụng kỹ thuật Synthetic Minority Oversampling Technique (SMOTE) giúp cải thiện độ chính xác trung bình thêm 7.51%, trong khi chuẩn hóa các từ lóng chỉ mang lại mức cải thiện khiêm tốn 0.25%. Ngược lại, việc đưa thêm lớp cảm xúc trung tính vào mô hình lại làm giảm độ chính xác xuống 9.67%. Cấu hình mô hình tối ưu được xác định là NBs kết hợp với SMOTE và chuẩn hóa từ lóng, đạt độ chính xác 93.53%, precision 93.65%, recall 93.53% và F1-score 93.47%. Phân tích đánh giá người dùng cũng cho thấy các chủ đề liên quan đến bảo mật, tính dễ sử dụng và tính kịp thời của ứng dụng chủ yếu nhận được phản hồi mang cảm xúc tiêu cực.

SENTIMENT ANALYSIS OF HEALTH INSURANCE APPLICATION REVIEWS USING NAÏVE BAYES, SUPPORT VECTOR MACHINE, AND SMOTE

Khau Van Bich [1], Vo Minh Tien[2], Tran Huu Duat[2*]

[1]*Tran Dai Nghia University, No. 189 Nguyen Oanh Str., 10 ward, Go Vap, Ho Chi Minh City, Viet Nam*
[2]*Thu Dau Mot University, No. 6, Tran Van On Street, Phu Loi Ward, Ho Chi Minh City, Vietnam*
*Corresponding author: *duatth@tdmu.edu.vn*

ARTICLE INFO

Received: Mar 6th, 2026
Revised: Mar 27th, 2026
Accepted: Apr 7th, 2026
Published: Apr 15th, 2026

KEYWORDS

Sentiment analysis;
Naïve Bayes;
Support Vector Machine;
SMOTE ;
Health Insurance Application.

ABSTRACT

This study explores factors shaping user experience in mobile health insurance application by analyzing user reviews through sentiment classification. Two models such as Naïve Bayes (NB) and Support Vector Machine (SVM) were examined, with NB yielding higher precision and F1-score. Data preparation proved influential: applying the Synthetic Minority Oversampling Technique (SMOTE) improved average accuracy by 7.51%, slang normalization added a modest 0.25%, while including neutral sentiment reduced accuracy by 9.67%. The selected configuration, NB combined with SMOTE and slang replacement, achieved 93.53% accuracy, 93.65% precision, 93.53% recall, and an F1-score of 93.47%. Examination of reviews further shows that discussions about security, ease of use, and timeliness are dominated by negative sentiment.

Doi: https://doi.org/10.61591/jslhu.26.1087
Available online at: https://lhj.vn

1. INTRODUCTION

Health is widely recognized as a fundamental social right that governments are expected to protect through public policy. One major initiative supporting this objective is the health insurance program, designed to expand access to healthcare services for the population [1]. Participation in this program has gradually become a requirement for accessing several public services [2]. To facilitate service delivery and reduce reliance on physical offices, a mobile-based health insurance application was introduced to enable participants to access services digitally. The platform provides several features, including new participant registration, personal data updates, digital membership cards, information services, and complaint submission [3].

However, increased service demand does not necessarily correspond with higher user satisfaction. Ratings displayed on the Google Play Store show a gradual decline for the health insurance application [4]. Adoption figures also reveal a gap between expectations and actual usage. This suggests that user perceptions of the platform require closer examination. Previous studies have attempted to identify factors influencing user experience through questionnaire-based surveys among application users [5]. Although useful, such approaches often require considerable time and resources to collect and process responses.

Sentiment analysis offers an alternative perspective by examining user reviews available on digital platforms as expressions of experience and opinion [6]. Machine learning models such as Naïve Bayes (NB) and Support Vector Machine (SVM) are commonly applied in this context due to their ability to produce reliable classification results with relatively efficient processing and limited training data [7]. Earlier studies have used these models to evaluate government service applications through reviews [8], analyze fintech services using Twitter data [9], and investigate public reactions during the COVID-19 pandemic [10]. However, datasets collected from online platforms frequently exhibit class imbalance, where certain sentiment categories dominate the data and reduce classification effectiveness. One widely used approach to address this issue is the Synthetic Minority Oversampling Technique (SMOTE), which balances the dataset by generating additional samples for underrepresented classes [11].

Particularly, the NB classifier is a probabilistic classification approach grounded in Bayes' theorem [12]. Its appeal lies in its simplicity; the model estimates class membership by calculating the probability distribution of features observed in the training data [13]. Because of this straightforward mechanism, NB has been widely applied in text-based tasks, particularly sentiment analysis, where classification decisions rely on the likelihood of word occurrences across categories [14].

Another commonly used method for classification and regression tasks is SVM. This algorithm applies the principle of structural risk minimization to determine an optimal hyperplane that separates data points belonging to different classes within the feature space [15]. In practice, SVM searches for the boundary that maximizes the margin between classes, allowing the model to generalize well when classifying unseen data.

During data collection, particularly when crawling online datasets, often exhibit class imbalance, a condition in which one category contains significantly more samples than others. Such an imbalance can negatively affect classification performance by biasing the model toward the dominant class. A widely adopted solution is the SMOTE, which addresses the imbalance by generating additional samples for minority classes through interpolation between neighboring data points [16]. This process helps balance the dataset and can improve the performance of sentiment classification models.

Beyond identifying sentiment polarity, digital service providers must also understand which aspects of their platforms shape user perception. Evaluating feedback through online service quality factors can highlight the service elements that require attention, enabling organizations to prioritize improvements and enhance overall service performance [17]. Service quality in digital environments is shaped by several interrelated elements, including the nature of interactions, the availability of information, and the effectiveness of internet-based services. These components, together with clearly defined service conditions, influence how users perceive reliability and trust in an organization’s digital platform. The researcher expanded the discussion of online service quality by examining its role in customer satisfaction within online brokerage services. Their findings outline several dimensions commonly used to evaluate service quality in digital systems.

These dimensions include responsiveness, referring to the readiness of a service provider to address user issues or obstacles; service reliability, which reflects the consistency and accuracy of delivered services; and ease of use, describing how intuitive the system interface and navigation are for users. Other dimensions involve competence, related to the system’s or provider’s ability to resolve problems effectively, and access, which concerns the availability of service channels and media through which users can interact with the platform. System reliability is also emphasized, focusing on technical stability and the absence of system failures or errors.

Timeliness represents another important factor, referring to system response speed and the regularity of updates. Finally, security addresses the protection of the system and user data, which plays a central role in maintaining trust in digital services.

Based on these considerations, this study analyzes factors influencing user experience in the health insurance application by combining sentiment analysis and online service quality factor evaluation. The research first compares the performance of NB and SVM classification models, incorporating SMOTE to address class imbalance. It then performs factor analysis to identify the most frequently discussed service aspects that require attention. The dataset used in this research consists of user reviews collected from the health insurance application on the Google Play Store. The findings are expected to provide insights that support future improvements to the application's service quality.

2. MATERIALS AND METHODS

The procedure followed a relatively direct trajectory: collecting user reviews, cleaning the text, classifying sentiment, and then identifying which service aspects appear most frequently. The analysis focuses specifically on user feedback related to the health insurance application, obtained from the review section of the Google Play Store. These reviews represent spontaneous user opinions posted publicly after interacting with the application, making them a practical source for understanding user experience issues without conducting surveys. The analytical framework combines sentiment classification with online service quality dimensions. Conceptually summarized in Figure 1, the workflow includes data collection, pre-processing, feature extraction, dataset division, labeling, class balancing using the SMOT, model construction, evaluation, sentiment interpretation, and factor analysis. The intention is not only to categorize sentiment but also to trace which service dimensions repeatedly emerge in user comments.

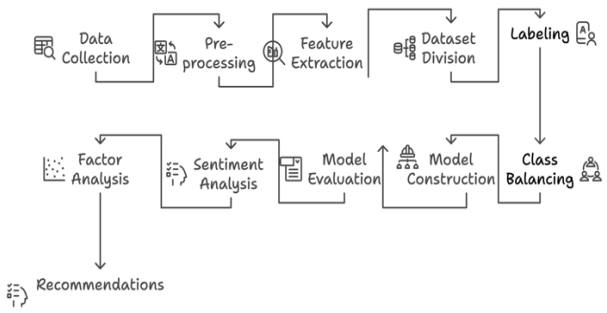

Figure 1: The workflow deployment

Data collection. The dataset was obtained through automated crawling of application reviews. A Python based tool, Google Play Scraper, was used to extract user feedback about the health insurance application from the Google Play Store. The collected dataset consists of approximately 3,000 user reviews posted within the observation period of the study. Each entry includes review text and associated metadata such as rating scores and timestamps. After extraction, the data were stored in structured spreadsheet formats: CSV or Excel, so they could be inspected, filtered, and prepared for further analysis. This approach allows the research to rely entirely on naturally occurring user feedback rather than responses generated through surveys or interviews.

Pre-processing. Before classification, the raw text required adjustment. Pre-processing addressed this stage. Reviews first underwent case folding, converting all characters to lowercase so identical words would not be counted separately. Informal expressions appeared frequently. Therefore, slang forms were replaced with their standard equivalents using a lexicon containing 15,396 entries. Tokenization followed, splitting sentences into individual tokens so each word could be handled computationally. Words carrying little semantic value: articles, connectors, common fillers, were removed through stop-word filtering using the Natural Language Toolkit. Finally, stemming reduced words with prefixes or suffixes to their root forms using Sastrawi. The outcome was simpler text, fewer redundant forms, and data more suitable for sentiment classification.

Feature extraction. After cleaning, the reviews were converted into numerical form through word-weighting. The representation used Term Frequency Inverse Document Frequency (TF-IDF), implemented through the scikit-learn environment in Python. Each document became a vector describing how often terms appear within a review relative to their distribution across the dataset.

Dataset division. The dataset was divided into several subsets. Training data served as learning material for the algorithms. Testing data evaluated model performance on unseen samples. Another subset remained for modeling and sentiment inference, where the selected classifier later assigned sentiment labels automatically.

Labeling. An internal team reviewed part of the dataset and assigned three sentiment categories: positive, negative, and neutral, to the training and testing data. These labeled instances became the reference during model learning. After training, the remaining review data were labeled automatically during sentiment classification.

Class balancing and classification model. Imbalanced sentiment distributions appeared early in the dataset, a

common situation with review data. To mitigate this issue, the study incorporated the SMOT, implemented through the imbalanced-learn library. The method generates synthetic samples for minority classes by interpolating between neighboring observations. Two classifiers were then examined: NBs and SVM. Both were implemented in scikit-learn. Training data were used to build the models, while testing data measured their performance.

Model evaluation relied on a confusion matrix generated through the scikit-learn evaluation utilities. Several experimental scenarios were conducted to observe how each classifier behaved under different configurations, allowing the study to determine the most reliable model before proceeding to sentiment interpretation and factor analysis.

Sentiment analysis. Once the strongest classifier was identified, it was applied to the remaining unlabeled review corpus. The algorithm assigned sentiment categories: positive, negative, or neutral, based on patterns learned during training. In practical terms, the model simply reads through the reviews and determines where each statement falls emotionally.

Factor analysis. After sentiment labelling, attention shifted to the content of the reviews themselves. The classified data were examined through the online service quality dimensions, with 8 dimensions guiding the grouping: responsiveness, service reliability, ease of use, competence, access, system reliability, timeliness, and security. Frequently occurring terms during classification were mapped to these dimensions. Words lacking clear relevance were excluded rather than forced into a category. The purpose was straightforward: identify which service aspects dominate user discussions and how those aspects relate to sentiment patterns.

Recommendations. The final stage translated analytical findings into practical suggestions. Factors associated with the highest concentration of negative sentiment were treated as priority areas for improvement in the health insurance application. Addressing these issues is expected to strengthen system performance and improve the overall user experience.

3. RESULTS AND DISCUSSION

3.1 Datasets

User review data were collected on December 30, 2024, from the review section of the health insurance application available on the Google Play Store. The extraction process used the Python library Google Play Scraper with the “most relevant” sorting filter, retrieving the latest 3,000 reviews. The dataset includes ratings from one to five stars, enabling the analysis to capture diverse user opinions. The crawling process was very efficient, taking approximately four seconds to obtain the full dataset. Compared with traditional survey-based approaches that may take several weeks to collect a relatively small number of responses, mining online reviews provides a faster and larger source of user feedback.

Before analysis, the raw review texts underwent several pre-processing steps. These included case folding to convert all text to lowercase, slang replacement to standardize informal expressions, tokenization to split sentences into individual words, stop-word removal to eliminate non-informative terms, and stemming to reduce words to their base forms. These steps transformed unstructured review sentences into normalized tokens, making them more suitable for computational analysis.

After pre-processing, the dataset was divided for model development. A total of 500 reviews were used as training data, while 100 reviews served as testing data to evaluate model performance. The remaining 2,400 reviews were reserved for automatic sentiment labeling using the selected classification algorithm. The training and testing subsets were manually annotated by the research team into three sentiment categories: positive, negative, and neutral, to support supervised learning and ensure reliable model training.

3.2 Performance of classification algorithms

This study compares the performance of two classification algorithms NB and SVM under several experimental configurations. Three additional variables were incorporated to examine their influence on model performance: slang word replacement, the use of the SMOTE algorithm for handling class imbalance, and the inclusion of neutral sentiment as a classification category. These configurations were expected to improve the effectiveness of the sentiment classification models. Model performance was evaluated using a confusion matrix to calculate accuracy, precision, recall, and F1-score. The detailed results of these experiments are summarized in Figure 2.

The results indicate that the best-performing configuration was obtained using the NBs algorithm combined with slang word replacement and the SMOTE algorithm, while excluding the neutral sentiment class. Under this configuration, the model achieved an accuracy of 93.53%, a precision of 93.65%, a recall of 93.53%, and an F1-score of 93.47%. These findings demonstrate that appropriate preprocessing and data balancing techniques can significantly enhance classification performance.

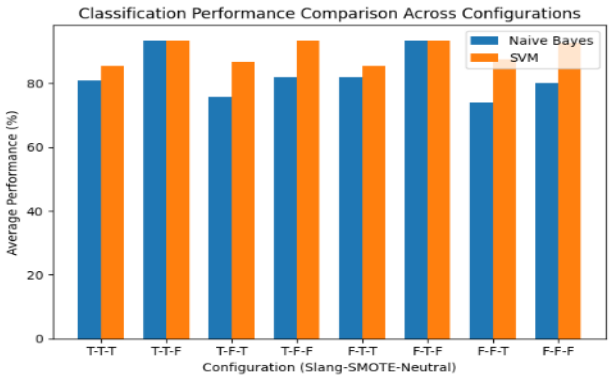

Figure 2. *The average performance NB vs SVM*

(T:True, F: False)

The results in Figure 3 show that integrating the SMOTE algorithm improved the average performance of both NB and SVM models by approximately 7.51%, increasing the overall score from 78% to 86.04%. Similarly, applying slang word replacement slightly improved the average performance by 0.25%, from 82.08% to 82.29%. In contrast, including neutral sentiment as a separate category reduced the overall performance by 9.67%, decreasing the average score from 87.08% to 77.29%. These findings suggest that data balancing techniques play a more substantial role in improving classification accuracy than text normalization techniques, while the addition of neutral sentiment introduces greater classification complexity that can reduce model performance.

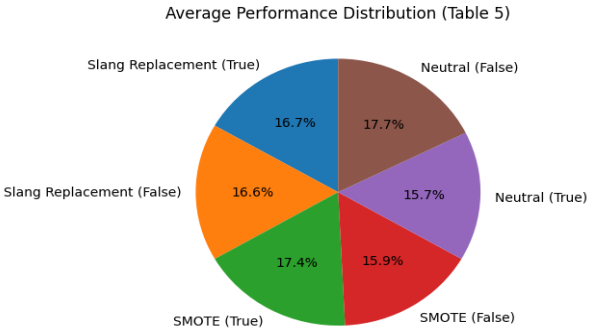

Figure 3. *The average performance distribution*

3.3 Analysis of sentiment results

Following sentiment classification, the study conducted a factor-based analysis to identify the aspects of service quality most frequently discussed by users. From the 3,000 reviews obtained from the Google Play Store**,** the most frequently occurring words were extracted. The 600 most frequent terms were then grouped according to the online service quality dimensions used in this research.

Both manually labeled data and automatically classified data were further analyzed using a service quality dictionary, allowing the identified terms to be associated with specific service quality factors. Through this process, it becomes possible to determine which aspects of the application most strongly influence user experience.

Table 1. *Sentiment analysis*

Category	Positive	Negative	Total
Responsiveness	20	105	125
Service reliability	187	634	821
Ease of use	446	1,509	1,955
Competence	224	237	461
Access	164	668	832
System reliability	53	481	534
Timeliness	289	1,105	1,394
Security	200	1,850	2,050

The factor analysis results are summarized in Table 1, which shows the distribution of positive and negative sentiments across eight service quality dimensions. The findings indicate that security is the most frequently discussed factor. A total of 2,050 reviews mention issues related to security, consisting of 200 positive reviews (9.75%) and 1,850 negative reviews (90.24%). This suggests that many users raise concerns about account safety, verification processes, or system protection. The second most discussed factor is ease of use, with 1,955 reviews, including 446 positive reviews (22.02%) and 1,509 negative reviews (77.98%). Many users commented on difficulties related to interface navigation, system complexity, or usability challenges. The third most prominent factor is timeliness, which appears in 1,394 reviews (46.03%), consisting of 289 positive reviews (20.20%) and 1,105 negative reviews (79.80%). These comments often relate to system response time, delays in service processing, or slow application performance. Overall, the factor analysis reveals that security, ease of use, and timeliness are the dominant dimensions influencing user perceptions of the health insurance application. The complete distribution of sentiment across all service quality factors is presented in Table 1, while the comparative influence of these factors is illustrated in Figure 4**.**

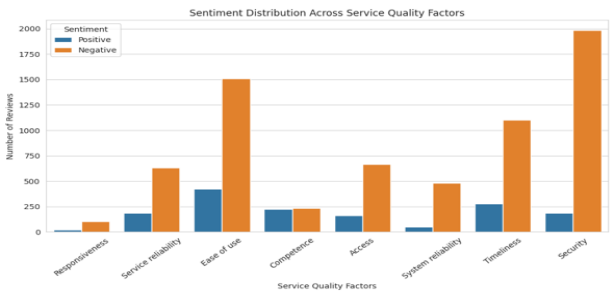

Figure 4. *The average performance distribution*

3.4 Recommendations

After identifying the three main critical factors, further research can focus on the most frequently discussed words related to these factors, as shown in Figure 5. By examining these frequently mentioned terms, researchers can better understand the causes of user dissatisfaction and develop strategies to improve service quality and user experience.

- In the security factor, commonly appearing words include *register, key, enter, login,* and *verification*, most of which are associated with negative sentiments. Therefore, the mobile health insurance application should simplify the registration and login process by providing additional options, such as integration with the Google Account API [18], and by benchmarking similar applications.
- For the ease-of-use factor, words such as *easy, difficult, good, menu,* and *hard* also tend to show negative sentiment. This suggests the need for further usability evaluation, improved feature layout, and user experience research using prototyping approaches before system development [19].
- In the timeliness factor, frequently mentioned words include *update, change, process,* and *direct*. The application should optimize the update schedule and consider technologies such as microservices and auto-scaling to improve application performance and responsiveness [20].

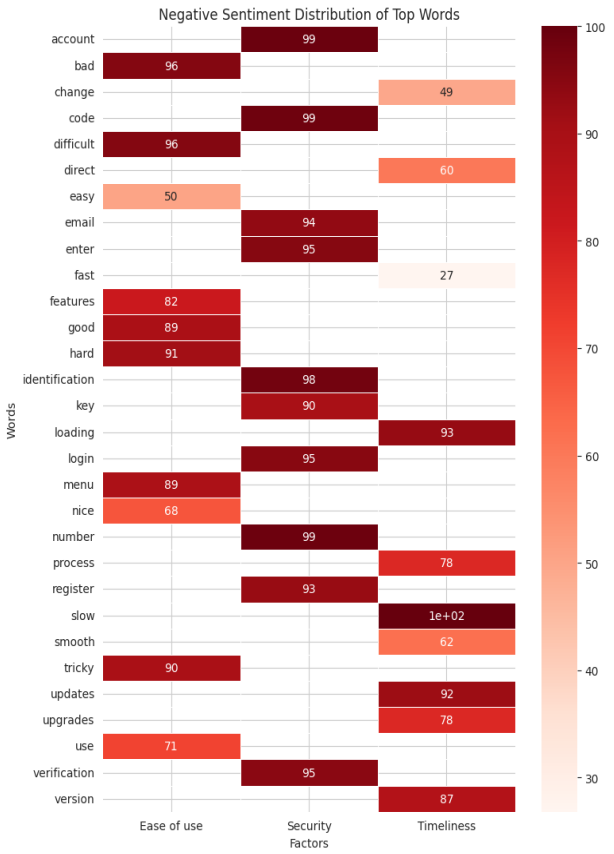

Figure 5. *The sentiment distribution top words*

4. CONCLUSION

This study aims to identify the factors that most influence user experience in the mobile health insurance application by applying an appropriate sentiment analysis model to user reviews. A comparison between the NB and SVM classifiers shows that NB performs better in terms of precision and F1-score. The integration of the SMOTE algorithm improved the average accuracy by 7.51%, while slang word normalization contributed a slight increase of 0.25%. In contrast, the inclusion of neutral data reduced the average accuracy by 9.67%. Based on these findings, the proposed model combines NB with SMOTE and slang word replacement, achieving an accuracy of 93.53%, precision of 93.65%, recall of 93.53%, and an F1-score of 93.47%. Further analysis of user reviews reveals that security, ease of use, and timeliness are the most critical factors affecting user satisfaction. Therefore, improvements in these areas are essential for enhancing the performance and overall user experience of the mobile application.

5. REFERENCES

[1] Djohanis, H., et al., (2025). Implementation Of The National Health Insurance Policy Within The Banggai District Health Office. Public Policy Journal, 6(1), 115-132.

[2] Ho, T. (2019). Optimizing healthcare policies through healthcare delivery and insurance design (Doctoral dissertation).

[3] Alawode, G. B., et al., (2025). Optimizing the health workforce for Universal Health Coverage: a framework for analysis and action. Human resources for health, 23(1), 27.

[4] Wisniewski, H., et al., (2019). Understanding the quality, effectiveness and attributes of top-rated smartphone health apps. BMJ Mental Health, 22(1).

[5] Nabasya, O. W., et al., (2024). Analysis of Service Quality Factors Affecting Patient Satisfaction in mHealth Teleconsultation. In 2024 Ninth International Conference on Informatics and Computing (ICIC) (pp. 1-6). IEEE.

[6] Zhu, L., et al., (2020). Sentiment and guest satisfaction with peer-to-peer accommodation: when are online ratings more trustworthy?. International Journal of Hospitality Management, 86, 102369.

[7] GhoshRoy, D., et al., (2022). Explainable AI to predict male fertility using extreme gradient boosting algorithm with SMOTE. Electronics, 12(1), 15.

[8] Hariyadi, H., et al., (2024). Implementasi Algoritma Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Ulasan Aplikasi Canva. Jurnal Minfo Polgan, 13(1), 261-269.

[9] Nurdin, T. A., et al., (2023). Sentiment analysis of user preference for old vs new fintech technology using

SVM and NB algorithms. Management Systems in Production Engineering, (4 (31), 373-380.

[10] Cahyono, H. D., et al., (2024). Fast Naïve Bayes classifiers for COVID-19 news in social networks. Indonesian Journal of Electrical Engineering and Computer Science, 34(2), 1033-1041.

[11] Chachoui, Y., et al., (2024). Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning. Computers and Education: Artificial Intelligence, 6, 100222.

[12] El Massari, H., et al., (2022). An ontological model based on machine learning for predicting breast cancer. International Journal of Advanced Computer Science and Applications, 13(7), 108-15.

[13] Apriliani, D., et al., (2020). Sentiment analysis for Indonesia hotel services review using optimized neural network. In Journal of Physics: Conference Series (Vol. 1538, No. 1, p. 012060). IOP Publishing.

[14] Antonius, R., et al., (2024). Pendekatan TF-IDF, SMOTE, dan SVM dalam Klasifikasi Sentimen Masyarakat terhadap Pemblokiran Judi Online. Buletin Ilmiah Informatika Teknologi, 2(3), 115-122.

[15] Ghaida, D. R., et al., (2024). SVM-Classified FaceNet and Eigenface Models under Lighting and Occlusion Variations for Face Recognition. In 2024 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 415-420). IEEE.

[16] Vu, V. H. (2024). Predict customer churn using combination deep learning networks model. Neural Computing and Applications, 36(9), 4867-4883.

[17] Siering, M., et al., (2018). Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. Decision support systems, 108, 1-12.

[18] Ryu, J., & Kim, T. (2025). Enhancing Hospital Data Security: A Blockchain-Based Protocol for Secure Information Sharing and Recovery. Electronics, 14(3), 580.

[19] Maduku, D. K., et al., (2023). Assessing customer passion, commitment, and word-of-mouth intentions in digital assistant usage: the moderating role of technology anxiety. Journal of Retailing and Consumer Services, 71, 103208.

[20] Taj, S., Daudpota, et al., (2025). Aspect-based sentiment analysis for software requirements elicitation using fine-tuned Bidirectional Encoder Representations from Transformers and Explainable Artificial Intelligence. Engineering Applications of Artificial Intelligence, 151, 110632.