Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 1, March 2025, pp. 201 – 209
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Natural Language Processing for Enhancing Anamnesis Documentation in Typhoid Fever Cases

[1]Tacyah Kholifah Putri, [2*]Mieke Nurmalasari, [3]Hosizah Hosizah,
[4]Dewi Krismawati, [5]Satria Bagus Panuntun
[1,2,3]Department of Health Information Management, Faculty of Health and Sciences, Universitas Esa Unggul
[4,5]Directorate of Statistical Analysis and Development, BPS Statistics Indonesia, Jakarta, Indonesia
Email: [1]ktacyah@gmail.com, [2]mieke@esaunggul.ac.id, [3]hosizah@esaunggul.ac.id,
[4]dewikrisma@bps.go.id, [5]satria.bagus@bps.go.id

Article Info
Article history:
Received Nov 4th, 2024
Revised Jan 8th, 2025
Accepted Feb 4th, 2025

Keywords:
Accuracy Value
Disease Anamnesis
Medical Records
Natural Language Processing
Typhoid Fever

ABSTRACT
The implementation of Natural Language Processing (NLP) is crucial for enhancing the quality of medical records. This study aimed to develop an NLP model to improve the accuracy of documenting disease anamnesis for typhoid fever. The problem addressed by this research is the difficulty in analyzing and classifying patient complaints recorded in electronic medical records, which can affect the accuracy of diagnosis and treatment. The urgency of this study lies in ensuring that documented medical information is used accurately to support diagnosis and patient management. A quantitative approach was used, focusing on electronic medical records of patients who underwent anti-salmonella IgM tests in 2023, involving 424 individuals. The study assessed the performance of three models: Support Vector Machines (SVM), Naive Bayes Bernoulli, and Logistic Regression. The SVM model achieved the highest accuracy at 81.4%, compared to 76.7% for Naive Bayes Bernoulli and 79.1% for Logistic Regression. Additionally, four topic models were identified, highlighting common complaint words and their impacts. The most frequently occurring symptoms in the anamnesis of typhoid fever were "defecation," "nausea," "vomiting," "fever," "diarrhea," "heartburn," "weakness," "loss of appetite," "abdominal pain," "cough," and "cold." This study demonstrates that the SVM model provides superior accuracy in analyzing medical records compared to other models.

Corresponding Author:
Mieke Nurmalasari.
Department of Health Information Management, Faculty of Health and Sciences, Universitas Esa Unggul
Jl. Arjuna Utara No.9, Duri Kepa, West Jakarta, Indonesia
Email: mieke@esaunggul.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i1.33325

1. INTRODUCTION

The development of digital technology has changed the paradigm in storing medical record data from the original manual to a more efficient electronic system. Electronic medical records are essential in modern health information systems, integrated with various subsystems to ensure adequate data storage and management [1].

Despite the significant advantages offered by Electronic Medical Records (EMR), challenges remain in optimizing their use, particularly in ensuring seamless integration across various healthcare service platforms. Several studies have highlighted issues with interoperability and data consistency across different systems [2][3]. For instance, Swarjana (2022) discussed biases in sampling techniques and data representation that hinder the effective use of EMR in diverse healthcare settings [2]. While improvements have been made, Ayumi et al. (2023) emphasized that existing EMR systems often fail to account for

unstructured data, such as patient narratives, which largely remain unaddressed in many healthcare environments [1].

Typhoid fever, caused by Salmonella Typhi, is a severe infection that can cause symptoms such as high fever, abdominal pain, and digestive problems. Without treatment, typhoid fever can lead to life-threatening complications. The disease usually spreads through contaminated food or water, making it a significant public health problem in many countries, especially in areas with poor sanitation [2].

Given the continuing prevalence of typhoid fever in developing regions, improving diagnostic and treatment protocols is critical to controlling outbreaks and reducing associated morbidity and mortality. Despite the advancements in Natural Language Processing (NLP) technology, there remains a need for further refinement in handling complex, domain-specific medical terminology to improve its application in clinical settingsolling outbreaks and reducing associated morbidity and mortality. Previous studies, such as the one by Ilyas (2020), have shown that early diagnosis through thorough anamnesis can significantly reduce complications [4]. However, in practice, incomplete or inconsistent anamnesis documentation remains a barrier to timely diagnosis. This issue is highlighted by studies like that of Rosady et al. (2022), which indicates how inconsistent patient history records often delay accurate diagnoses, particularly in resource-limited settings [5].

NLP is a rapidly developing branch of computer science, especially in its application to understanding and processing human language [3]. In medical records, NLP has excellent potential to improve the quality of patient history recording, often in the form of accessible narratives, so that the information contained therein can be organized and used more effectively, especially in the medical diagnosis process [6]. NLP focuses on developing a deeper understanding of language by moving from syntax-based text manipulation, such as word counting, to natural language processing that takes into account semantic, grammatical, and contextual constraints [5].

Despite the advancements in NLP technology, there remains a need for further refinement in how it handles complex, domain-specific medical terminology to improve its application in clinical settings. Several NLP-based models, such as those proposed by Rachman (2021), have proven effective in analyzing structured data but often struggle with the intricacies of medical language [6][7]. Badawi (2021) noted that there is a gap in fully automating NLP tools to process unstructured patient data, especially in multilingual or diverse healthcare systems [3].

The anamnesis process, which is the initial stage in the interaction between patients and medical personnel, plays a vital role in collecting relevant information to establish a diagnosis of the disease. Anamnesis is the process of collecting medical information from patients through interviews, including medical history, symptoms, and factors that may affect their health condition [3]. The quality of anamnesis records is essential in determining the accuracy of the diagnosis and subsequent treatment. However, challenges often arise due to the lack of standardization and consistent structure in recording anamnesis [5].

The variability in how anamnesis is recorded across different healthcare providers underscores the need for standardized methods to ensure more accurate and reliable data capture. As shown by Ilyas (2020), inconsistent data quality in anamnesis forms can directly affect the diagnostic accuracy of conditions like typhoid fever [4]. Several researchers have explored standardizing anamnesis records with NLP [8], but their methodologies often rely on manual data preparation, which can be time-consuming and prone to errors.

Previous research has shown that the application of Natural Language Processing can overcome this problem by changing unstructured anamnesis text into more structured information, which can be used to support clinical decisions [8]. This is important considering the findings that writing a less structured anamnesis can complicate the final diagnosis process, as occurred in several cases in the hospital.

Building upon these findings, there is an opportunity to implement NLP more comprehensively, addressing data structure and language complexity challenges, ultimately leading to better-informed clinical decisions. Unlike previous studies, this research aims to automate the extraction of clinical features from anamnesis text, making the process more efficient and reducing human error. This study distinguishes itself by using advanced NLP techniques that do not require manual intervention, ensuring a more scalable and reliable solution.

2. RESEARCH METHOD

This study uses an observational approach by directly examining the medical records of inpatients undergoing anti-salmonella IgM tests. Data were collected through observation sheets, which included anamnesis, supporting examination results, and positive or negative typhoid fever diagnoses. The research method involves several operational stages, described in more detail in the following sub-sections.

2.1. Dataset

Text preprocessing is the preparation or initial processing of text data before further analysis. in this stage, data were collected from medical records that included anamnesis (patient complaints), supporting examination results, and final diagnosis (positive or negative for typhoid fever). The dataset consists of 424 inpatient records, all undergoing anti-salmonella IgM testing. This data was used to analyze the relationship between the patient's symptoms and the final diagnosis of typhoid fever.

2.2. Text Preprocessing

Before further analysis, text preprocessing is carried out to prepare the data for modeling. This involves several steps: (1) Case Folding: This step converts all uppercase letters in the text to lowercase to ensure uniformity. (2) Tokenization: This step breaks the text into smaller units called tokens, which can be words, phrases, or characters. (3) Filtering: Irrelevant elements such as punctuation, numbers, and common words (stopwords) are removed from the text. (4) Stemming: This process reduces words to their root form by removing suffixes, which helps reduce morphological variation and simplify word representation. This preprocessing step is crucial for cleaning the data and ensuring it is in a suitable format for the subsequent analysis.

2.3. Modelling: Classification Analysis

After preprocessing, classification analysis is conducted to differentiate between positive and negative typhoid fever cases based on the collected anamnesis data. Three machine learning models were applied in this analysis: (1) Support Vector Machine (SVM): This model works by finding the best hyperplane that separates data points between two classes. The goal is to maximize the margin between the two classes. (2) Naive Bayes Bernoulli: This probabilistic model is used for text classification by assuming feature independence and using the Bernoulli probability model. (3) Logistic Regression: A linear model used for predicting the probability of a given text belonging to a particular category. Each model was evaluated using accuracy as the performance metric, calculated as the ratio of correct predictions to the total number of predictions.

2.4. Data Visualization: Topic Modeling

After classification, data visualization is performed using Topic Modeling to analyze the positive typhoid fever disease anamnesis from 2023. Latent Dirichlet Allocation (LDA) was used for topic modeling, identifying topics frequently appearing in the patient's anamnesis. The pyLDAvis tool was employed to visualize the topics, allowing an in-depth understanding of the relationships between topics and words in the data. Topic modeling helps uncover common symptoms and patterns associated with typhoid fever diagnosis, contributing valuable insights into the clinical characteristics of the disease.

2.5. Natural Language Processing Model: Word Cloud

A word cloud was created to illustrate the frequency of symptoms reported during anamnesis. In this visualization, the size of each word corresponds to its frequency, with larger words indicating more commonly mentioned symptoms. This approach provides a quick overview of the most prevalent symptoms of typhoid fever, such as fever, nausea, vomiting, diarrhea, and abdominal pain.

2.6. Mathematical Models and Equations

Several mathematical models were employed in this study to perform the analysis, and the corresponding equations are presented as follows:

1. Support Vector Machine (SVM): The goal of SVM is to find the optimal hyperplane that maximizes the margin between two classes. The equation for the hyperplane is equation 1.

$$w \,.\, x + b = 0 \tag{1}$$

where w is the weight vector, x is the feature vector, and b is the bias term.

2. Naive Bayes Bernoulli: The probability of a class C given a set of features $X = \{x_i\,, x_2, \ldots\, x_n$ is computed as equation 2.

$$P(C|X) = \frac{P(C)\pi_{I=1}^{n} P(x_i\,|C)}{P\,(X)} \tag{2}$$

where $P(C)$ is the prior probability of the class, and $P(x_i|C)$ is the likelihood of feature x_i given the class C.

3. Logistic Regression: The probability of class y=1y = 1y=1 is calculated using the logistic function as equation 3.

$$P(y-1|x) = \frac{1}{1+e^{-\beta_0+\beta_{1x1}+\beta_{2x2}+\ldots+\beta_{nxn}}} \quad (3)$$

where $\beta_0, \beta_1, \ldots, \beta_n$ are the model coefficients, and $x_1, x_2, \ldots, x_n$ are the input features.

2.7. Overview of Methodology

Figure 1 presents an overview of the entire research methodology, including dataset collection, text preprocessing, classification analysis, topic modeling, word cloud generation, and the mathematical models used in the analysis.

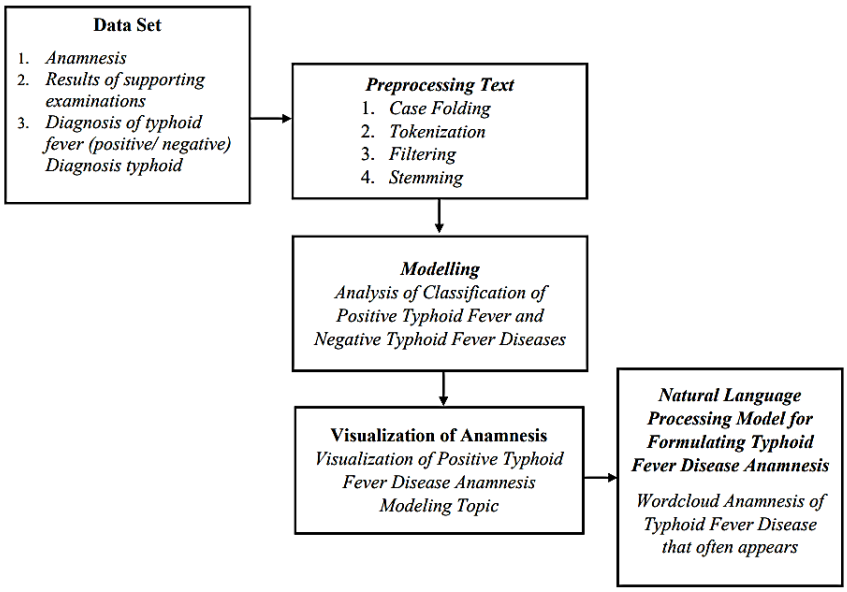

Figure 1. Research Stages

3. RESULTS AND ANALYSIS

In this study, the modelling stage was carried out with text preprocessing, classification analysis, displaying topic modelling and word clouds.

3.1. Text Preprocessing in NLP for Typhoid Fever Analysis

Text preprocessing is a preparation or initial processing process on text data before further analysis or modelling [9]. Before evaluating model accuracy, a preprocessing stage is carried out to observe how it affects model performance, especially in accuracy. The preprocessing stage is carried out in the following manner: (1) Case folding is the first step in text preprocessing that changes all uppercase characters in the text to lowercase [10], aiming to match the text format without considering the capitalization of letters. (2) Tokenization is an important process in text preprocessing that breaks text into smaller units, called tokens, which can be words, phrases, or characters, depending on the text's complexity [11]. (3) Filtering is a step in text preprocessing that aims to eliminate parts of the text that are irrelevant or do not significantly contribute to the analysis, such as punctuation, numbers, symbols, and common words [12]. (4) Stemming is a process in text preprocessing that changes words to their basic form or root form by removing word endings, thereby reducing morphological variation and producing a more straightforward representation of words with the same root or meaning [13].

Effectively performing these preprocessing steps ensures the text data is well-prepared, leading to more accurate and relevant results in subsequent analysis or modeling. Preprocessing plays a critical role in medical text analysis by reducing noise in the data, enabling machine learning models to concentrate on the most pertinent information. Given the variations often present in medical terminology, preprocessing standardizes the text, enhancing the model's ability to recognize patterns. This process significantly improves the quality of the text data and strengthens the model’s capacity to interpret and understand the content

3.2. Natural Language Processing Modeling for Typhoid Fever Anamnesis

Modelling for the formulation of typhoid fever anamnesis was done through classification and prediction analysis between positive and negative diagnosis cases. This analysis evaluated three machine learning models: Support Vector Machines (SVM), Naive Bayes Bernoulli, and Logistic Regression. The parameter compared in this analysis is accuracy, which is a measure to assess the model's performance. Accuracy is the ratio between the number of correct predictions and the total number of predictions made. In other words, accuracy indicates how well the model classifies data into the correct category. For example, if the model correctly predicts 80 out of 100 data, its accuracy is 80%. The data used in this analysis is 424 medical records of inpatients who underwent anti-salmonella IgM examination, consisting of positive and negative diagnosis cases. The accuracy results for each algorithm evaluated are in Table 1.

The results demonstrate that the SVM model significantly outperforms Naive Bayes Bernoulli and Logistic Regression, showcasing its superior ability to accurately classify typhoid fever cases. This advantage is likely attributed to SVM's capability to handle complex, high-dimensional data—a characteristic often encountered in medical text analysis. However, while SVM excels in accuracy, its computational demands can become prohibitive when applied to larger datasets. Future research could explore optimization techniques or ensemble methods to achieve a balance between accuracy and computational efficiency. Conversely, Naive Bayes Bernoulli, despite its slightly lower accuracy, remains a strong contender for real-time applications due to its computational simplicity and efficiency, particularly in resource-constrained settings. These findings align with studies such as Ayumi et al. (2023), which also reported SVM's superior performance in disease prediction tasks involving text classification. Nevertheless, simpler models like Naive Bayes have demonstrated acceptable performance in specific contexts, especially when training data is limited, as highlighted by Ilyas (2020). The differences in accuracy across the models underscore the importance of selecting an algorithm that best suits the data's complexity while balancing accuracy, interpretability, and computational cost.

Table 1. Accuracy Results

Algorithm	Accuracy (%)
Support Vector Machine (SVM)	81.4%
Naive Bayes Bernoulli	76.7%
Logistic Regression.	79.1%

Based on the accuracy results shown in Table 1, the SVM model achieved the highest accuracy at 81.4%, followed by Logistic Regression at 79.1%, and Naive Bayes Bernoulli at 76.7%. This indicates that the SVM model performs better than the other two models in accurately classifying the typhoid fever cases in the dataset.

SVM works by finding a hyperplane that separates data between classes, with the margin defined as the distance from the hyperplane to the closest data point from each class [14]. This may explain its superior performance, as SVM is particularly effective in high-dimensional spaces, such as medical text data. In contrast, Naive Bayes Bernoulli assumes feature independence and is based on the Bernoulli probability model. Despite its lower accuracy, this model is easy to implement and often works well for text classification tasks [15]. Logistic Regression, which is also a linear model, showed slightly better performance than Naive Bayes Bernoulli but still lagged behind SVM in terms of accuracy [16].

The findings of this study align with previous research that highlights SVM’s superior performance in medical text analysis compared to other models. For instance, Ilyas (2020) demonstrated that SVM outperformed Naive Bayes and Logistic Regression in classifying medical records. However, research like Ayumi et al. (2023) has shown that, while less accurate, Naive Bayes models can be valuable in scenarios where model interpretability is a priority.

These results reinforce the importance of selecting an appropriate model based on the task's specific requirements, such as accuracy, interpretability, and computational efficiency. Future studies could investigate ensemble methods to combine the strengths of different models, potentially enhancing overall performance while maintaining a balance between prediction accuracy and interpretability.

3.3. Visualization of Typhoid Fever Anamnesis

Visualization of anamnesis of positive typhoid fever disease that often appears in 2023 using topic modeling pyLDAvis from positive typhoid fever diagnosis anamnesis data. pyLDAvis is a tool used to visualize and understand topic models generated using the Latent Dirichlet Allocation (LDA) algorithm [17]. This tool allows users to explore relationships between topics, relevant words, and topic distributions within documents [18]. It provides valuable insights into text data, helping healthcare professionals to identify

patterns that may not be immediately evident in standard analysis. The visualization in Figure 2 shows the topic modeling results for positive typhoid fever diagnoses observed in 2023, leveraging pyLDAvis.

The pyLDAvis topic modeling results give a more detailed view of the common symptoms associated with typhoid fever. This helps healthcare professionals identify the symptoms most frequently reported by patients.

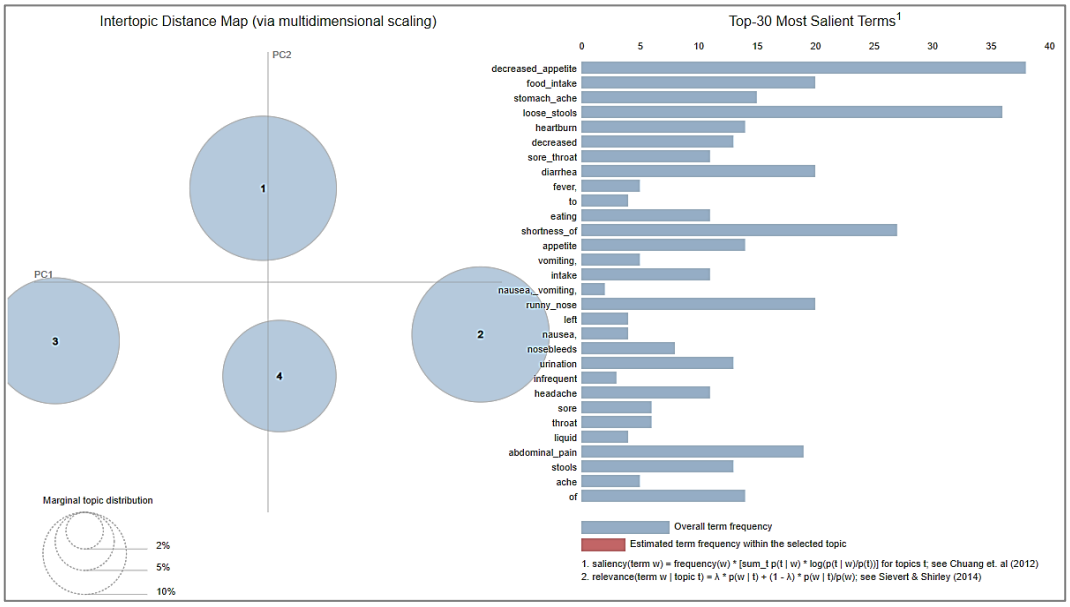

Figure 2. Visualization of Typhoid Fever Disease Anamnesis

The results of the topic modeling from Figure 2 produced four distinct topics, as shown in Table 2. The analysis reveals common symptoms associated with positive typhoid fever cases, including gastrointestinal issues like loose stools and abdominal pain, respiratory symptoms such as cough and sore throat, systemic manifestations like weakness and dizziness, and inflammatory symptoms such as fever and vomiting. This comprehensive view highlights the diverse clinical presentation of the disease.

Table 2. Topic Modeling Word Results

Topics	Say
Topic 1 (Gastrointestinal Symptoms)	loose_stools, decreased_appetite, shirtness_of, heartburn, runny_nose, diarrhea, weakness, cough, of, shortness, breath, node, runny, abdominal_pain, stools, headache, loose, appetite, decreased, stomach_ache, defecation, gums, blood, bleeding_gums, bleeding, liquid, pain, body, urination, sometimes
Topic 2 (Respiratory and General Symptoms)	loose_stools, food_intake, abdominal_pain, shortness_of, stools, loose, sore_throat, nosebleeds, pain, blood, gums, urination, of, phlegm, shortness, breath, food, intake, cough, left, abdominal, infrequement, sore, throat, weak, eating, chest, red, nosebleed, hours
Topic 3 (Systemic Manifestations)	decreased_appetite, stomach_ache, diarrhea, decreased, appetite, eating, runny_nose, sore_throat, weakness, weak, headache, to, stomach, ache, runny, throat, cough, nose, drinking, abdominal_pain, blood, dizziness, sore, body, urination, shortness_of, loose_stools, less, dry, pain
Topic 4 (Inflammatory and Infectious Symptoms)	decreased_appetite, shortness_of, loose_stools, food_intake, urination, cough, intake, weakness, of, shortness, breath, headache, runny_nose, nosebleed, fever, pain, abdominal_pain, decreased, vomiting, appetite, blood, loose, nausea, food, nausea_vomiting, diarrhea, runny, stools, urinating, nose

3.4. Natural Language Processing Model for Analyzing Typhoid Fever Anamnesis

Natural language processing models use wordclouds. A wordcloud is a visualization of text where the size of the words indicates the frequency or relevance of the words in the text. Words that appear more frequently are displayed more prominent, helping to illustrate key themes in the document [19]. The purpose of a wordcloud is to present a clear visualization of word frequency in a text, allowing users to identify key themes or topics quickly[20]. Figure 3 illustrates the wordcloud representing the anamnesis of typhoid fever patients, highlighting the terms that frequently appear.

The word cloud in Figure 3 reflects the typical clinical presentation of typhoid fever, with terms like "nausea," "vomiting," and "fever" standing out as the most frequently mentioned. This reinforces the primary symptoms patients and clinicians focus on during the initial assessment. Visualizing word frequency with a word cloud aids clinicians in quickly identifying key symptom patterns, providing an additional tool for prioritizing diagnostic considerations. The prominence of gastrointestinal and systemic symptoms in the word cloud aligns with the established clinical characteristics of typhoid fever, validating the utility of NLP for processing complex medical data to support diagnosis. Word clouds are particularly effective in clinical decision support systems, enabling doctors to prioritize the most indicative symptoms of typhoid fever for further investigation.

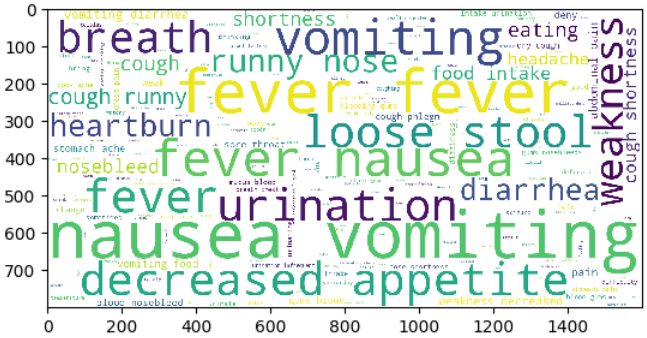

Figure 3. Wordcloud Positive Typhoid Fever History

The word cloud illustrates the most frequently mentioned words in the anamnesis of typhoid fever patients. Prominent words include "nausea," "vomiting," "fever," "diarrhea," "weakness," "headache," and "abdominal pain." The size of these words reflects their frequent appearance in the records, highlighting key symptoms.

This study demonstrates that NLP and machine learning models can effectively classify typhoid fever cases, with SVM showing the highest accuracy. Using pyLDAvis for topic modeling provides insightful visualizations that enhance understanding of the common symptoms associated with the disease. However, there are limitations, including the relatively small dataset of 424 records and potential biases due to variations in recording anamnesis. Additionally, more computational resources are needed for models like SVM when applied to larger datasets. Future studies should explore ensemble models to combine the strengths of multiple algorithms and improve performance. Moreover, expanding the dataset to include more diverse patient records could further enhance the generalizability of the model and its clinical.

4. CONCLUSION

The results of this study indicate that the Support Vector Machines (SVM) model outperforms other methods in classifying typhoid fever anamnesis, achieving the highest accuracy of 81.4%, compared to 76.7% for the Naive Bayes Bernoulli model and 79.1% for Logistic Regression. This demonstrates the effectiveness of SVM in predicting typhoid fever diagnoses based on patient anamnesis. Additionally, the topic analysis reveals four main themes featuring key terms commonly associated with typhoid fever, such as "nausea," "vomiting," "fever," "decreased appetite," "diarrhea," "shortness of breath," "urination," "heartburn," "headache," "weakness," and "cough." The generated word cloud highlights these complaints as focal points in diagnosing typhoid fever. This study significantly contributes to the advancement of Natural Language Processing (NLP) models in formulating typhoid fever anamnesis, improving diagnostic accuracy and disease management. By transforming unstructured anamnesis data into structured information, the system is expected to assist clinicians in making more accurate and efficient medical decisions.

However, this study has several limitations, including the relatively small dataset (424 medical records), which may impact the model's generalizability. Additionally, inconsistencies in how anamnesis is documented could introduce biases into the data. To address these limitations, future research should focus on expanding the dataset to include a broader range of medical records and employing model optimization techniques or ensemble approaches to enhance performance. Further studies could also investigate more advanced NLP methods or extend the analysis to cover other infectious diseases, thereby increasing the model's applicability and effectiveness in clinical settings. Overall, the findings from this study highlight the significant potential of NLP applications in healthcare, particularly in modeling and analyzing anamnesis data for infectious diseases like typhoid fever. With further development, this technology can become a valuable tool for supporting diagnosis and disease management in the future.

REFERENCES

[1] V. Ayumi, H. Noprisson, M. Utami, E. D. Putra, and M. Purba, *Basic Concepts of Natural Language Processing (NLP)*, 1st ed. West Java: CV Jejak, Anggota IKAPI, 2023.

[2] I. K. Swarjana, *Population-Sample, Sampling Techniques and Bias in Research*, 1st ed. Yogyakarta: Andi (Member of IKAPI), 2022.

[3] A. Badawi, 'The Effectiveness of Natural Language Processing (NLP) As a Processing Solution and Semantic Improvement', *Int. J. Econ. Technol. Soc. Sci.*, vol. 2, no. 1, p. 42, 2021, doi: 10.53695/injects.v2i1.194.

[4] A. A. Ilyas, 'The Correlation Between The Completeness of Patient Anamnesis form and External Causes Diagnosis Code Accuracy in Bahagia Hospital Makassar', *Aptirmiki*, vol. 5, 2020.

[5] D. Septriana Rosady, L. Lazuardi, and S. Sastrowijoto, 'Clinical Teleconsultation: Ethics, Discipline, and Medical Law', *Indones. Heal. Law J.*, vol. 2, no. 01, pp. 1–23, 2022, doi: 10.53337/jhki.v2i01.17.

[6] C. I. Ratnasari, S. Kusumadewi, and L. Rosita, 'Natural Language Processing Model for Formulating Patient Complaints', *Natl. Semin. Med. Informatics V*, vol. 3, no. 1, p. 17, 2014.

[7] F. P. Rachman, 'Comparison of Deep Learning Models for Sentiment Analysis Classification with Natural Language Processing Techniques', *J. Inf. Technol. Manag.*, vol. 7, no. 2, pp. 113–121, 2021, doi: 10.26905/jtmi.v7i2.6506.

[8] A. R. Lubis and M. K. Nasution, 'Twitter Data Analysis And Text Normalization In Collecting Standard Word', vol. 4, no. 2, p. 858, 2023.

[9] J. E. Widyaya and S. Budi, 'The Effect of Preprocessing on Diabetic Retinopathy Classification with the Transfer Learning Convolutional Neural Network Approach', *J. Informatics Eng. Inf. Syst.*, vol. 7, no. 1, p. 112, 2021, doi: 10.28932/jutisi.v7i1.3327.

[10] E. B. Susanto, P. A. Christianto, M. R. Maulana, and S. W. Binabar, 'Performance Analysis of Naïve Bayes Algorithm on Community Sentiment Dataset of NEWSAKPOLE Samsat Central Java Application', *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 3, pp. 234–241, 2022, doi: 10.37859/coscitech.v3i3.4343.

[11] Y. Akbar and T. Sugiharto, 'Sentiment Analysis of Twitter Users in Indonesia Towards ChatGPT Using C4.5 and Naive Bayes Algorithms', *J. Sci. Technol.*, vol. 5, no. 1, p. 117, 2023.

[12] P. Salsabiila, 'Analysis of Public Sentiment on Twitter Regarding the Johnny Depp and Amber Heard Defamation Trial Case Pravidya', *Indones. J. Humanit. Soc. Sci.*, vol. 5, no. 1, pp. 401–416, 2024.

[13] A. E. Budiman and A. Widjaja, 'Analysis of the Influence of Text Preprocessing on Plagiarism Detection in Final Project Documents', *J. Informatics Eng. Inf. Syst.*, vol. 6, no. 3, p. 476, 2020, doi: 10.28932/jutisi.v6i3.2892.

[14] H. N. Irmanda and R. Astriratma, 'Classification of Pantun Types Using the Support Vector Machines (SVM) Method', *J. Syst. Eng. Inf. Technol.*, vol. 4, no. 5, p. 917, 2020, doi: 10.29207/resti.v4i5.2313.

[15] H. Azizah, B. S. Rintyarna, and T. A. Cahyanto, 'Sentiment Analysis to Measure Public Trust in the Procurement of Covid-19 Vaccines Based on Bernoulli Naive Bayes', *J. Inf. Technol. Comput. Eng.*, vol. 3, no. 1, pp. 23–29, 2022, doi: 10.37148/bios.v3i1.36.

[16] D. Y. Utami, Uni, E. Nurlelah, and F. N. Hasan, 'Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to predict diabetes', *J. Informatics Telecommun. Eng.*, vol. 5, no. 1, pp. 53–64, 2021, doi: 10.31289/jite.v5i1.5201.

[17] A. Ariansyah and U. Indahyanti, 'Feature Extraction in Topic Modeling Using Latent Dirichlet Allocation Method in Data Leakage Events', *Indones. J. Appl. Technol.*, no. 2, pp. 1–24, 2024.

[18] R. P. Sardika, C. Asyraq, M. R. Pribadi, and W. Widhiarso, 'ChatGPT Topic Modeling in Twitter Reviews Using the Latent Dirichlet Allocation Method', *J. Inf. Syst. Inf. Technol. Comput.*, vol. 14, no. 2, pp. 80–149, 2024, [Online]. Available: enosalmungg@gmail.com

[19] I. T. Julianto and Lindawati, 'Sentiment Analysis of the Garut Institute of Technology Academic Information System', *J. Algorithms*, vol. 19, no. 1, pp. 449–456, 2022, doi: 10.33364/algoritma/v.19-1.1112.

[20] R. R. S. Putri Kumala Sari, 'Comparison of Support Vector Machine and Random Forest Algorithms for Metaverse Sentiment Analysis', *J. Mnemon.*, vol. 7, no. 1, pp. 31–39, 2024.

BIBLIOGRAPHY OF AUTHORS

Tacyah Kholifah Putri graduated from Health Information Management Depatment, Universitas Esa Unggul. She works as a Medical Recorder and Health Information at Sumber Waras Hospital as Head of the Medical Records Installation.

Mieke Nurmalasari, she is a lecturer in the Health Information Management Department, Faculty of Health Sciences, at Universitas Esa Unggul. Her expertise is statistics, biostatistics, statistical modelling, and research interests in data mining.

Hosizah, currently working as a lecturer in the Health Information Management Department, Faculty of Health Sciences. Her expertise includes health information management, with a focus on electronic health records, telemedicine, medical data mining, and application of health information systems.

Dewi Krismawati, she works at the Directorate of Statistical Analysis and Development, BPS Statistics Indonesia. She is a statistician who is passionate about information technology. Currently, she is very excited to learn big data technology, data engineering, and data science.

Satria Bagus Panuntun, he is a Data Engineer at the Directorate of Statistical Analysis and Development, Badan Pusat Statistik (BPS Statistics Indonesia).