



Original Research

LDA-SVM-based topic modeling for healthcare document classification and medical decision support in low-resource settings

ID Etana Fikadu Dinsa*

Department of Computer Science, Institute of Technology, Wollega University,
Nekemte, Oromia, Ethiopia

Abstract

There is great potential and many challenges in the increased number of texts related to the field of healthcare, especially in low-resource languages like Afaan Oromo. The proper classification and retrieval of the text documents is very important in making a medical decision. This paper provides a hybrid approach using Latent Dirichlet Allocation (LDA) for interpretation of topics and Support Vector Machine for proper classification of the documents. The use of the hybrid approach on the healthcare documents will enable us to have access to relevant information in linguistically poor areas. We used the EHR dataset in this study and applied different pre-processing methods to it. A feature extraction technique was utilized. In this case, the developed model LDA-SVM has shown good performance, achieving an accuracy of 93.67% and clear interpretation of topics. This model has been successful in classifying EHRs, especially in dealing with unlabeled datasets. But in this research, only a monolingual model was developed for the Afaan Oromo language. Expanding this work to support multilingual, including code-mixed and cross-lingual data, remains an important direction for future research.

Copyright © 2026 STAR Journal, Wollega University. All Rights Reserved.

Article Information

Article History:

Received: 26-05-2026

Revised: 30-08-2026

Accepted: 30-09-2026

Keywords:

EHRs, Topic Modeling,
LDA, SVM,
Classification, Medical
Decision Support

*Corresponding
Author:

Etana Fikadu Dinsa

E-mail:

etanaf@wollegauniversity.edu.et

INTRODUCTION

Through digitalization of healthcare systems, large quantities of unstructured documents such as diagnostic reports, clinical notes, and discharge summaries are available (Wu et al., 2021). Such reports contain rich medical information and hold enormous potential for enhancing clinical workflow, research, and public health interventions (Dinsa et al., 2024a). However, their effective use remains challenging, especially in settings where language resources, computational resources, and trained operators are limited (Fikadu, 2020; Stewart et al., 2023; Dinsa et al., 2026; Fikadu et al., 2025). Resource-scarce language like Afaan Oromo is greatly

underrepresented in EHR natural language processing (NLP) studies. The unavailability of annotated corpora, language-dependent tools (Fikadu et al., 2019), and uniform medical vocabularies is delaying the development of intelligent healthcare applications for these languages (Dinsa, 2024b; Negash Tesso et al., 2023). Such digital inequality leads to unequal access to medical results and decision-support tools, especially for rural and low-income patients. Conventional rule-based or keyword-based information retrieval methods fail to capture the contextual richness and variability of medical vocabulary, particularly in the case of translation

or localization into low-resource languages. New, scalable, and interpretable approaches need to be formulated to enable effective document analysis in such environments. The Latent Dirichlet Allocation algorithm presents a statistical approach to finding latent semantic structures in text data such that interpretable topic models can be found (Muthusami et al., 2024; Cahyo et al., 2024). When integrated with Support Vector Machines, a high-performance classifier, the resulting hybrid model can both uncover informative patterns and classify documents with accuracy (Thielmann et al., 2023). This integration ensures that the classification is not only efficient but also understandable an essential requirement for medical doctors and policy-makers.

This paper focuses on how the LDA-SVM model can be reformed and applied in the healthcare domain in the context of Afaan Oromo documents. Most existing studies rely on English or other resource-rich languages, leaving underrepresented languages like Afaan Oromo overlooked. Few efforts have attempted to localize topic models or machine learning pipelines for Afaan Oromo. A notable challenge in these settings is the lack of labeled data and linguistic tools, which hampers direct adoption of conventional NLP methods. In this research, we demonstrate how LDA-SVM can improve access to and usage of critical medical information, contribute to equitable digital health innovation, and provide groundwork for future work in low-resource settings. This study addresses these limitations by customizing an LDA-SVM pipeline tailored for healthcare document classification in Afaan Oromo. It contributes to bridging the gap in bilingual healthcare NLP by providing a replicable framework for low-resource language settings.

Statement of the problem

The ever-increasing amount of unstructured medical text in the form of clinical notes, discharge summaries, and scientific literature poses many difficulties in making timely and precise decisions in healthcare, especially in poor countries where annotated corpora, computational

Sci. Technol. Arts Res. J., July –Sep, 2026, 15(3), 129-138
resources, and domain knowledge are few. Traditional methods of document classification require a large labeled dataset, which is difficult to gather in such conditions. There is, therefore, a demand for techniques that can use latent semantic structure from very little data. Combining Latent Dirichlet Allocation (LDA), a model of topics in texts, with Support Vector Machines (SVM), a classifier, is an attractive solution for organizing medical documents and hence supporting decision-making in evidence-based medicine while still being possible in under-resourced health care environments.

Previous studies have explored a range of machine learning techniques for clinical document classification, with a significant focus on high-resource languages such as English (Farkhod et al., 2021; Hu et al., 2014). Topic modeling methods like LDA have been widely used in biomedical informatics to identify latent themes in large document collections. Blei et al.(2003) introduced LDA as a generative probabilistic model capable of discovering abstract topics that occur in a corpus. In the healthcare domain, LDA has been applied to tasks such as automatic summarization of clinical notes, clustering patient records, and identifying risk factors for diseases.

Thielmann et al. (2023) illustrate the mixture of SVM and LDA topic modeling as a multi-classification approach that avoids the issues of human labeling. The results indicate that unsupervised textual article classification conducted via web-scraping, SVM, and LDA modeling provides very accurate classification outcomes for diverse data sets. But these methods also need the domain expert's support to determine the LDA output topics retrieved. The work (Yang & Zhang, 2018) discovers text extraction, mining, and examining helpful information from unstructured text documents using sentiment analysis and LDA-topic modeling by analyzing English Twitter text documents. These techniques allow individuals to discover data more effectively and efficiently.

The work (Dinsa et al., 2024a) proposed a topic modeling approach using LDA to address

the challenge of unstructured health-related documents in Afaan Oromo. All collected documents are written in Afaan Oromo and lack label information, which poses substantial challenges for classifying the documents and applying supervised learning approaches. Based on the prepared dataset, the model provides promised values. The current work is an extension of this paper by developing a hybrid model using SVM classification algorithms.

Research questions

1. How powerfully can an LDA–SVM framework classify medical documents in low-resource settings?
2. To what extent do topic representations generated by LDA improve the accuracy and robustness of medical document classification for healthcare decision support?
3. What are the practical challenges and deployment considerations for implementing an LDA–SVM-based medical decision support system in resource-constrained healthcare environments?

MATERIALS AND METHODS

Dataset

Data acquisition is the most problematic activity in this study area, especially in resource-scarce languages. In our case, we used health records data, which were in Afaan Oromo, to build our research corpus. Our target dataset consists of 3000 health-related documents collected from public health social media. All the documents collected do not contain label information, and hence it is hard to label or classify them perfectly. This deficiency of labeled data poses serious challenges in the structuring of the documents and in the implementation of supervised learning algorithms that are based on labeled datasets. The dataset we used in this experiment is collected from disease symptoms and disease-related

Sci. Technol. Arts Res. J., July –Sep, 2026, 15(3), 129-138
information in the health sector, as well as publicly available social media data. The data consist of information shared online as public messages, for example, communications from health authorities to the community. Pre-processing tasks involve cleaning and transforming raw medical questions, such as removing stop words, lowercasing, and tokenizing to prepare the text for effective model training.

Text Pre-processing

First, the obtained set of health-related documents was processed using various text pre-processing procedures to ensure the improvement of input data quality and consistency. In particular, unnecessary symbols, including punctuation marks, numbers, URLs, and other unrelated elements, were removed from the data, and then text normalization as well as case transformation were performed. Next, tokens were extracted from the data, and commonly used stop words that do not convey much semantic information were deleted. When needed, stemming was performed to transform the words to their stem or root form. The pre-processed text corpus was further used as input for Latent Dirichlet Allocation (LDA) to find latent topics in order to represent each document in terms of a topic distribution vector, which was further used as input features for classifiers including Logistic Regression and SVM.

Proposed methodology

The proposed methodology utilizes LDA for generating labeled documents through latent-topic extraction and then trains classification algorithms for further classification of unobserved documents. Our method architecture can be separated into five (5) diverse modules. These components can be presented in the [Figure 1](#). The modules are: 1. Data Pre-processing and Dictionary creation, 2. Topic modeling using LDA, 3. Model output, 4. Classification model and 5. Classification model evaluation

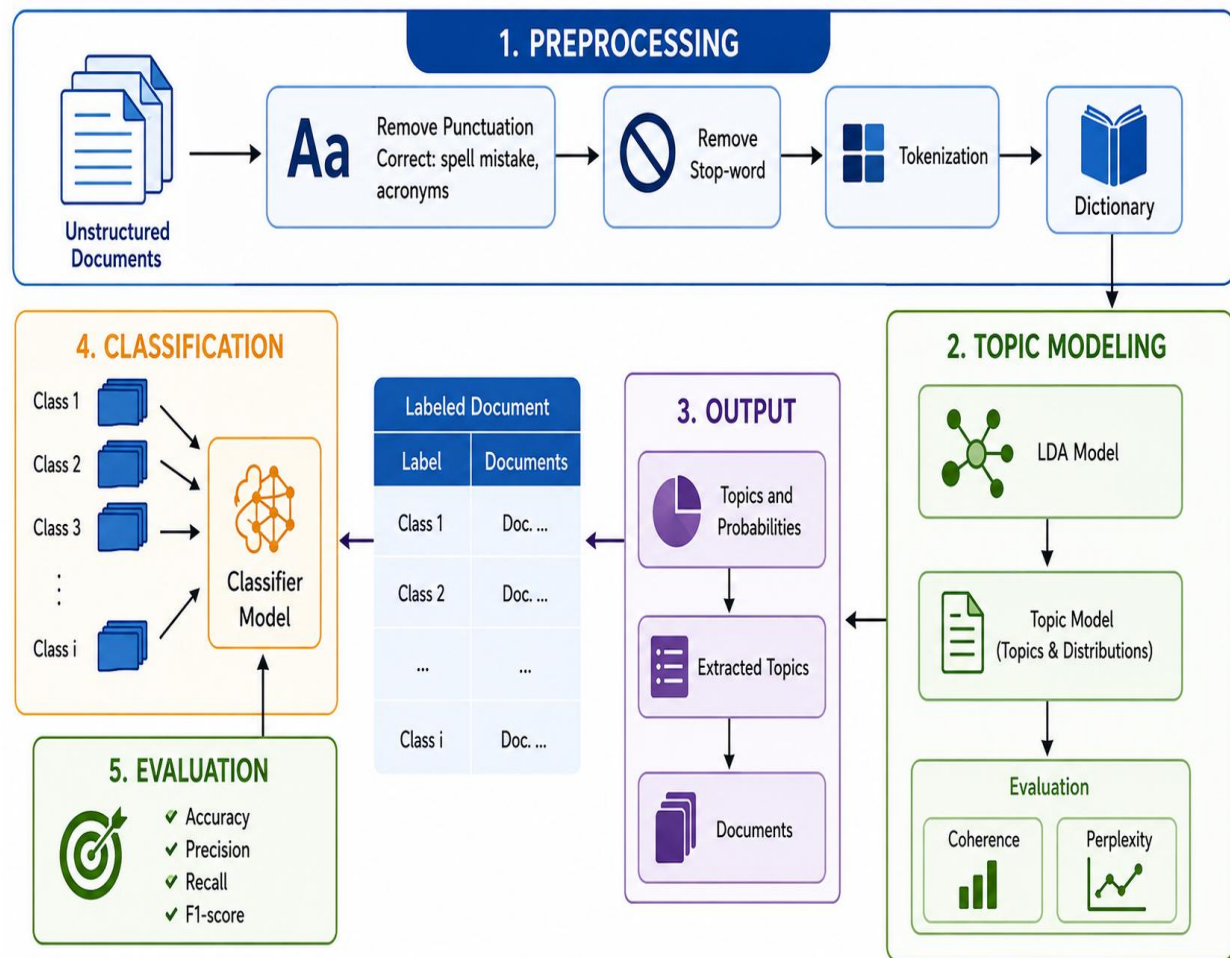


Figure 1. High-level Architecture of the Proposed Methodology

Topic Modeling with LDA

LDA was applied to discover the hidden thematic structures within the collections of records. It assumes that each document is a combination of numerous topics, and each theme is characterized by a distribution over words. To decide the optimal number of topics, we evaluated multiple models with different topic counts using coherence scoring, a measure that assesses the semantic similarity of high-probability words in a topic. In topic modeling, coherence value refers to the similarity measure for each word in the topics identified (Rogers & Longo, 2017). High topic coherence values are obtained for interpretable topics when that happens. Different measures using the co-occurrence score of the key words in each separate topic have been proposed for use in the topic mining arena. The best way to find out if

a topic is good is by measuring its coherence. Following the training process, each document was represented by a vector representing the distribution of topics in that document, where each element of the vector represented the probability of that document being associated with a particular topic. The extracted topic vectors served as interpretable, low-dimensional features for downstream classification. This work builds upon Table 2 of Dinsa et al. (2024a), which illustrates some of the topics extracted from electronic health documents, which form the basis for our proposed methodology. Based on the experiment, each topic can be observed to contain 10 words corresponding to the topic. For each word gathered under each topic, there will be a respective probability ratio. We used the 10 most probable words per topic based on the optimum

coherence value. So, the LDA model was implemented using scikit-learn with 10 topics. The document-topic prior (α) and topic-word prior (η/β) were set to their default values of 0.1 ($1/K$, where $K = 10$). The model was trained using batch variational Bayes with a maximum of 10 iterations, a mean-change convergence tolerance of 0.001, and a perplexity tolerance of 0.1. Based on the experimental results, the extracted topics were assigned labels according to their keywords and distributions to facilitate interpretation.

Machine learning classifier

The experimentation was carried out using Python programming on Windows -10 -Pro -64 bit, and

Sci. Technol. Arts Res. J., July –Sep, 2026, 15(3), 129-138 the following hardware specifications: Intel (R)-Xeon(R)-Silver-4210R CPU-@-2.40, GHz 2.39GHz, 32 GB of installed memory (RAM), and 3 TB of hard disk drive. As we have a trained LDA model on unstructured documents to extract latent topics and associated documents, we wanted to go further with the classification algorithms (Chauhan & Shah, 2022). We have conducted a document classification experiment according to the results of our LDA model, which is a labeled dataset. In this study, we proposed that support vector machines (SVM) be used to achieve classification tasks as illustrated in Figure 2.

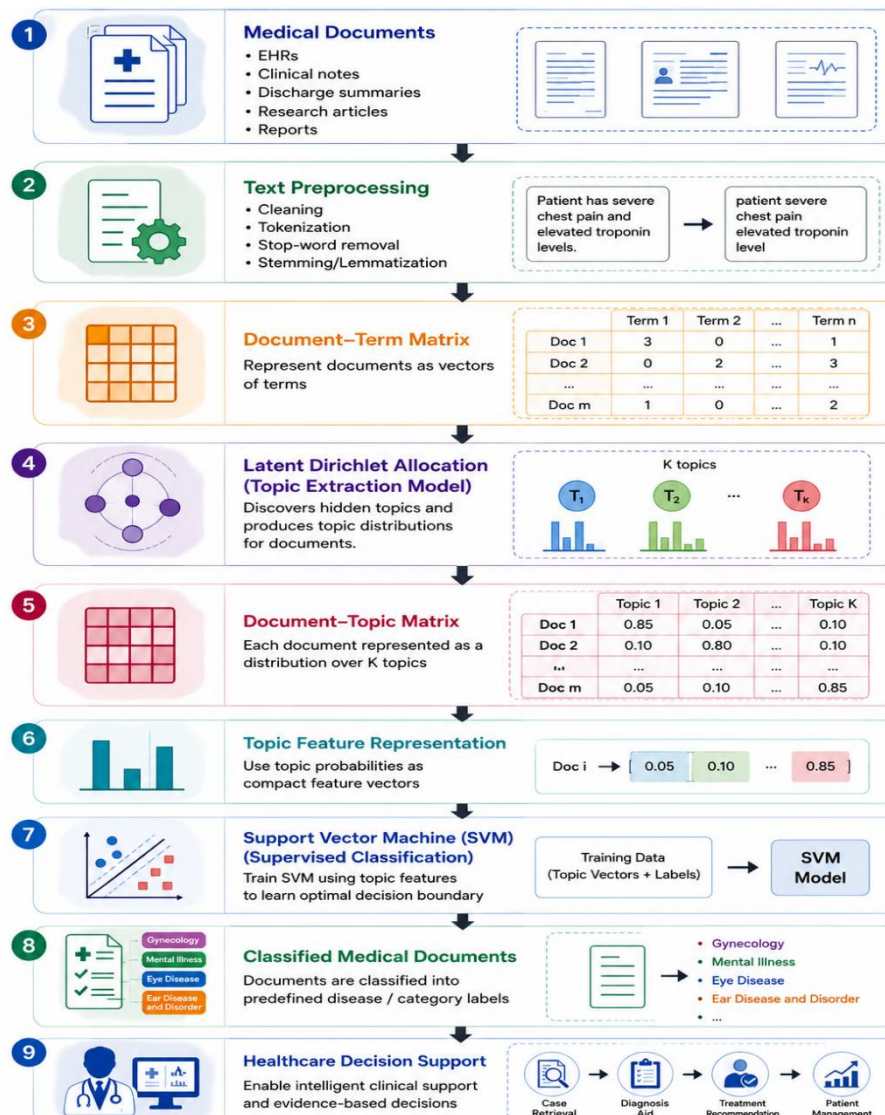


Figure 2. General framework of proposed model (LDA-SVM)

SVM is a powerful statistical machine learning technique that was first introduced by the author (Lu, 2013). SVM was extensively used in text document classification tasks due to its capability to handle lots of data inputs and its good performance. In these experimental results, the output of LDA, which means extracted topics and topic probabilities or distributions, is crucial data for the classifier. At this point, the topics and topic probabilities are used as input to the SVM, which would result in labeled topics that would finally be useful to classify unseen documents. In this work, for comparison of performance of the new model, Logistic Regression and Random Forest were employed. The results show that the new LDA-SVM model outperforms others with respect to greater accuracy and generalizability on unseen data by virtue of its ability to learn underlying topic patterns and apply strong classification boundaries. The topic distribution vectors generated by the LDA model were used as input features for a SVM classifier. To assess the classifier's performance, we used the standard metrics such as accuracy, recall, precision, and F-measure (Fikadu et al., 2025; Sumathi & Raja, 2023).

RESULTS AND DISCUSSION

Results

We conducted classification experiments with the SVM, Logistic Regression, and Random Forest. Each model is assessed using the K-fold cross-validation technique (Shakeel et al., 2018). We used k=5 in this cross-validation to evaluate the generalization error of the predictive models. In k-

Sci. Technol. Arts Res. J., July –Sep, 2026, 15(3), 129-138

fold cross-validation, the training data was separated into k equal-sized subsets. The result of the k-fold cross-validation is the average of the performances gained from the k rounds. For all the classification outcomes, we used the static parameter settings of this operator: average performances only= true, number of validations = 5, sampling type (to build the subsets) = stratified-sampling. We iteratively train these models on subsets of data and validate against the held-out data. We have used SVM as the main ML algorithm for training the classifier. As we observed from the experimental results, the accuracy obtained from SVM is 93.67%, Multinomial Logistic Regression is 91.83%, and Random Forest is 92.67%. The logistic regression model results in a massive variance, which indicates that the model is overfitting to its training data compared to the other models. Logistic regression can therefore form unstable linear decision boundaries that are sensitive to fold-specific samples, whereas SVM maximizes the separating margin and is generally less sensitive to such variation. Random forest has a tiny edge with a good median accuracy over logistic regression and fewer SVM models. However, it has been discovered that there is an improvement in the overall classification accuracy of support vector machines with LDA (LDA+SVM), with an accuracy of 93.67% and an F1 score of 93.65%.

Table 1 and Figure 3 present the results analysis of model comparison with the LR, SVM, and RF models. The accuracy, recall, precision, and F1 score are illustrated clearly.

Table 1

Experimental results of classifiers.

Item	Classifier	Accuracy	Precision	Recall	F1 Score
LDA	SVM	93.67	93.85	93.70	93.65
	RF	92.67	92.98	92.67	92.67
	LR	91.83	92.02	91.83	91.77

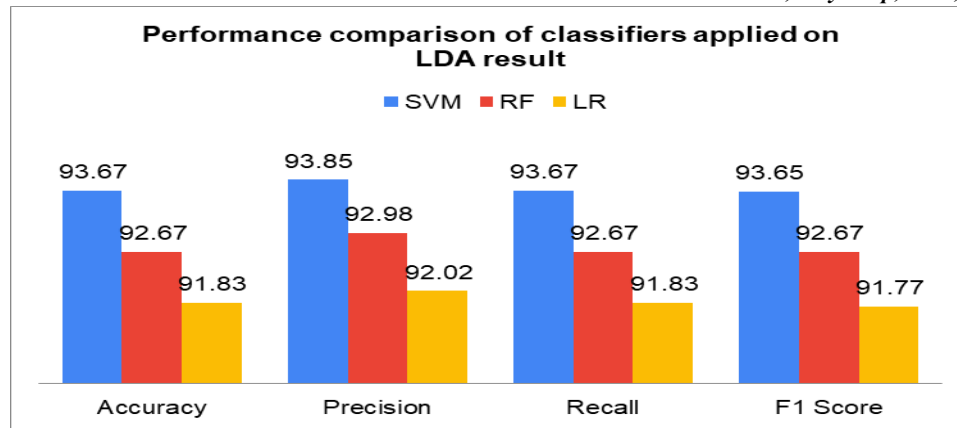


Figure 3. Graphical representations of classifiers results

Discussion

In this article, it was discussed how one can discover hidden topics in unstructured health documents and convert them into a useful form for information-retrieval and classification tasks that are applied in many important processes to achieve the goals of healthcare organizations. As mentioned earlier, the LDA (Latent Dirichlet Allocation) algorithm is applied in this research paper; it is a widely used technique in topic modeling and can help discover hidden topics in a collection of records. The approach is very simple and easy to interpret. LDA model is language-independent and works using the idea of a bag-of-words model where the order of words does not matter. A document is a combination of topics, and a topic is a combination of words, so it becomes easy to understand the results of LDA.

Conceptual Application of LDA-SVM for Information Retrieval and Decision Support

The collaboration between LDA and SVM is a powerful model for both information retrieval (IR) and recommendation systems in the electronic

healthcare context. By exploiting LDA's capability to generate understandable topic representations, healthcare documents can be indexed and retrieved based on semantic relevance rather than keyword matching. This facilitates more accurate and informative access to information by clinicians and decision-makers. Once latent subjects are accessed, the SVM classifier avoids misclassifying documents so that user query-based personalized or case-specific recommendations can be provided. Together, the LDA-SVM pipeline enhances clinical decision support by delivering precise, context-dependent information based on individualized medical needs, which is most critical in low-resource settings where time, resources, and computing resources are limited. The user is expected to provide the query and the number of documents required to be retrieved, and the system will return the relevant documents based on the top similarity value. Figure 4 illustrates that the top documents based on similarity and relevance to the user query are retrieved.

Please enter your question (or type 'exit' to quit):
garaa keessa na uursa, na ciniina

Query	relevant document	Similarity Probability
garaa keessa na uursa, na ciniina	garaa keessa uursa ciniina mataa ulfaata	0.8121
garaa keessa na uursa, na ciniina	garaa keessaa uursa ciniina mataa kos ulfaata	0.7058
garaa keessa na uursa, na ciniina	garaa keessaa uursa ciniina mataa kos ulfaata	0.7058
garaa keessa na uursa, na ciniina	gurra keessa uursa	0.7006
garaa keessa na uursa, na ciniina	gurra keessa uursa	0.7006
garaa keessa na uursa, na ciniina	olna deebisa gubanyaata dhowwa garaa keessa uursa	0.6557
garaa keessa na uursa, na ciniina	olna deebisa gubanyaata dhowwa garaa keessa uursa	0.6557
garaa keessa na uursa, na ciniina	gurra keessa boombii uursa	0.6363
garaa keessa na uursa, na ciniina	garaatu ciniina	0.6135
garaa keessa na uursa, na ciniina	garaatu ciniina	0.6135

Similarity Probability	Number of Documents
0.8121	1
0.7058	2
0.7006	2
0.6557	2
0.6363	1
0.6135	2

Please enter your question (or type 'exit' to quit):
exit
Exiting...

Figure 4. The evidence retrieved from the corpus based on the user's query

Practical Challenges

The process of implementing the proposed LDA-SVM model as a medical decision support system in a resource-limited healthcare setting demands a careful approach to the modeling process. One of the main difficulties is the ability to collect a sufficient amount of high-quality, labeled medical text data that might be insufficient, imbalanced, contain noise, or be written in a low-resource language. Effective text pre-processing, including data cleansing, tokenization, removing stop words, and performing stemming or lemmatization, is required to obtain valuable topics via LDA. The right choice of the number of topics, proper LDA hyperparameter tuning (e.g., choosing values for α and β), and SVM parameter tuning (e.g., selecting kernel type, C and gamma) is necessary for providing good classification results without overfitting. Moreover, feature selection and dimensionality reduction should be considered to increase computational efficiency on equipment with low computational resources. In terms of deployment, the system should be light, scalable, interpretable, and able to cope with medical knowledge evolution via retraining on a regular basis. Inclusion of domain knowledge, while building the model and developing an easy-to-use interface for healthcare specialists, is an additional component that makes such a decision support system effective.

CONCLUSION

This paper introduces a hybrid LDA-SVM approach tailored for optimizing information retrieval and electronic healthcare document classification in low-resource environments for Afaan Oromo. By combining explainability with classification capability, the trained model provides an efficient answer to the advancement of digital health tools and accessibility for under-represented linguistic populations. The LDA-SVM model performs extremely well on this dataset, demonstrating its effectiveness at modeling underlying topic distributions and improving classification performance. Its success is a good sign of its capability to handle challenging, domain-specific text in low-resource settings.

Recommendations

Future work will address the following areas in an attempt to improve even further the performance, scalability, and application of the suggested LDA-SVM-based medical text classification system.

Future investigations will consider the incorporation of advanced feature representation schemes, such as contextual word embeddings and transformer-based language models, to represent complex semantic relations in the medical text.

In the future, the work will involve building large and diverse sets of Afaan Oromo medical texts and linguistic resources that could be used to

Etana Fikadu Dinsa

tackle problems related to language processing in low-resource languages.

We suggest that the LDA–SVM scheme will be improved using hybrid modeling that combines topic modeling with deep neural architectures to improve disease classification and generalization ability.

The investigation will be continued in terms of optimizing the suggested model parameters, feature selection, and class imbalance handling to increase the robustness of the classification algorithm. Finally, the developed algorithm will be tested in practice by implementing the approach using the technique of explainable artificial intelligence in the context of real healthcare institutions.

CRedit authorship contribution statement

Etana Fikadu Dinsa: Conceptualization, Methodology, Data Preparation, Validation, writing code, experiments, Investigation, Writing an original draft, and review and editing.

Competing Interests

The authors state that they have no conflict of interest.

Ethical Approval

Not applicable

Availability of supporting data

The datasets used during the current study will be available from the corresponding author upon request.

Acknowledgement

The author thanks the Department of Computer Science, Institute of Technology, Wollega University, Nekemte, Ethiopia, for providing the necessary support to complete this study.

REFERENCES

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3, 993–1022.
<https://dl.acm.org/doi/10.5555/944919.944937>

Cahyo, P. W., Aesy, U. S., & Santosa, B. D. (2024). Topic Sentiment Using Logistic

Sci. Technol. Arts Res. J., July–Sep, 2026, 15(3), 129-138

Regression and Latent Dirichlet Allocation as a Customer Satisfaction Analysis Model. *Jurnal Infotel*, 16(1), 1–16.

<https://doi.org/10.20895/infotel.v16i1.1081>

Chauhan, U., & Shah, A. (2021). Topic Modeling Using Latent Dirichlet allocation. *ACM Computing Surveys*, 54(7), 1–35.

<https://doi.org/10.1145/3462478>

Dinsa, E. F., Das, M., & Abebe, T. U. (2024a). A topic modeling approach for analyzing and categorizing electronic healthcare documents in Afaan Oromo without label information. *Scientific Reports*, 14(1), 1–14.

<https://doi.org/10.1038/s41598-024-83743-3>

Dinsa, E. F., Das, M., & Abebe, T. U. (2024b). AI-based disease category prediction model using symptoms from low-resource Ethiopian language: Afaan Oromo text. *Scientific Reports*, 14(1), 11233.

<https://doi.org/10.1038/s41598-024-62278-7>

Dinsa, E. F., Das, M., & Abebe, T. U. (2026). Advances in electronic health records enabled by artificial intelligence and natural language processing: a review of recent developments, limitations and future applications. *Discover Applied Sciences*, 8(3), 250.

<https://doi.org/10.1007/s42452-026-08318-5>

Dinsa, E. F., Das, M., Abebe, T. U., & Ramaswamy, K. (2024). Automatic categorization of medical documents in Afaan Oromo using ensemble machine learning techniques. *Discover Applied Sciences*, 6(11), 575.

<https://doi.org/10.1007/s42452-024-06307-0>

Dinsa, E. F., & P, R. B. (2019). Application of data mining Classification Algorithms for AFAAN Oromo Media text news Categorization. *International Journal of Computer Trends and Technology*, 67(7), 73–79.

<https://doi.org/10.14445/22312803/ijett-v672>

Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021). LDA-Based Topic Modeling Sentiment Analysis using Topic/Document/Sentence (TDS) model. *Applied Sciences*, 11(23), 11091.

<https://doi.org/10.3390/app112311091>

A Peer-reviewed Official International Journal of Wollega University, Ethiopia

- Fikadu, E. (2020). Comparative study of the Performance of Machine Learning Text Classifiers Applied to Afaan Oromo Text. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, 77–83. <https://doi.org/10.32628/cseit20645>
- Fikadu, E., Das, M., Urgessa, T., & Ramaswamy, K. (2025). A supervised learning approach for recommending medical specialists in the healthcare sector for the Afaan Oromo context. *Discover Computing*, 28(1), 46. <https://doi.org/10.1007/s10791-025-09556-8>
- Fikadu, E., Urgessa, T., & Das, M. (2025). MLP-SVM: a hybrid approach for improving the performance of the classification model for health-related documents from social media using multi-layer perceptron and support vector machine. *Discover Applied Sciences*, 7(4), 320. <https://doi.org/10.1007/s42452-025-06851-3>
- Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning*, 95(3), 423–469. <https://doi.org/10.1007/s10994-013-5413-0>
- Lu, Y. (2013). Automatic topic identification of health-related messages in online health community using text classification. *SpringerPlus*, 2(1), 1–7. <https://doi.org/10.1186/2193-1801-2-309>
- Muthusami, R., Mani Kandan, N., Saritha, K., Narenthiran, B., Nagaprasad, N., & Ramaswamy, K. (2024). Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. *Scientific Reports*, 14(1), 12003. <https://doi.org/10.1038/s41598-024-61738-4>
- Negash Tesso, D., Fikadu Dinsa, E., & Fikadu Kenani, H. (2023). Signed Language Translation into Afaan Oromo Text Using Deep-Learning Approach. *American Journal of Artificial Intelligence*, 7(2), 40–51. <https://doi.org/10.11648/j.ajai.20230702.12>
- Sci. Technol. Arts Res. J., July –Sep, 2026, 15(3), 129-138
- Rogers, N., & Longo, L. (2017). A comparison on the classification of short-text documents using Latent Dirichlet Allocation and Formal Concept Analysis. *CEUR Workshop Proceedings*, 2086(December), 50–62. https://ceur-ws.org/Vol-2086/AICS2017_per_2_7.pdf
- Shakeel, K., Tahir, G. R., Tehseen, I., & Ali, M. (2018). A framework of Urdu topic modeling using latent dirichlet allocation (LDA). 2018 *IEEE 8th Annual Computing and Communication Workshop and Conference*, CCWC 2018, 2018-Janua(June 2020), 117–123. <https://doi.org/10.1109/CCWC.2018.8301655>
- Stewart, J., Lu, J., Goudie, A., Arendts, G., Meka, S. A., Freeman, S., Walker, K., Sprivulis, P., Sanfilippo, F., Bennamoun, M., & Dwivedi, G. (2023). Applications of natural language processing at emergency department triage: A narrative review. *PLoS ONE*, 18(12), 1–25. <https://doi.org/10.1371/journal.pone.0279953>
- Sumathi, M., & Raja, S. P. (2023). Machine learning algorithm-based spam detection in social networks. *Social Network Analysis and Mining*, 13(1), 1–13. <https://doi.org/10.1007/s13278-023-01108-6>
- Thielmann, A., Weisser, C., Krenz, A., & Säfken, B. (2023). Unsupervised document classification integrating web scraping, one-class SVM and LDA topic modelling. *Journal of Applied Statistics*, 50(3), 574–591. <https://doi.org/10.1080/02664763.2021.19190>
- Wu, M., Du, X., Gu, R., & Wei, J. (2021). Artificial Intelligence for Clinical Decision Support in Sepsis. *Frontiers in Medicine*, 8(5), 1–9. <https://doi.org/10.3389/fmed.2021.665464>
- Yang, S. and Zhang, H. (2018). Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis. *International Journal of Computer and Information Engineering*, 12(7), 525–529. <https://doi.org/10.5281/zenodo.1317350>