Advancing Sentiment Analysis with LDA Topic Modelling and VADER.

Bharathi Mohan G
Department. of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
g_bharathimohan@ch.amrita.edu

Aparna S
Department. of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
aparnasureshkumar24@gmail.com

Kothamasu Rishi
Department. of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
kothamsurishi2@gmail.com

Maddineni Sravani
Department. of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
sravanimaddineni24@gmail.com

***Abstract*—This paper compares sentiment analysis techniques using Latent Dirichlet Allocation (LDA) topic modelling and traditional sentiment analysis methods. Sentiment analysis is important for understanding the sentiment expressed in collected data such as customer reviews and social media posts. Our study evaluates the performance of sentiment analysis with and without topic modelling using key evaluation metrics, including accuracy, precision, recall, and F1-score. Experimental results obtained from analyzing a dataset of textual reviews demonstrate the effectiveness of incorporating LDA topic modelling into sentiment analysis pipelines. The sentiment analysis model with topic modelling achieves an accuracy of 72.46%, outperforming the accuracy of 71.17% obtained by the traditional sentiment analysis method. Similarly, prototypes also show improvements in accuracy, recall, and F1 scores, with precision at 72.73%, recall at 72.7%, and F1-score at 72.53%, compared to 71.4%, 72.25%, and 71.7% respectively, for sentiment analysis without topic modelling. These findings highlight the potential benefits of integrating LDA topic modelling into sentiment analysis frameworks, providing in-depth insight into the views expressed in the data files and enhance perspective on the overall analysis of various applications.**

Keywords—Sentiment Analysis, Topic Modelling, IMDb review, Polarity VADER, LDA, Coherence Score.

I. Introduction

Sentiment analysis is the process of extracting, analyzing and classifying various emotions expressed in text using vocabulary and word processing. It is well known that when choosing what to see or watch next moviegoers referred to as users or reviewers for the remainder of this paper typically consult ratings and reviews for the writers of this paper this is in fact the case. Sentiment analysis, also called sentiment mining, is a process of automatically extracting or classifying the positive, negative and neutral feelings of the reviewers. However, extracting sentiment polarity or sentiment mining from multiple data sources is a difficult process. For instance, the ability to handle complex sentences is not sufficient and the polarity identification and accuracy of the current Sentiment Analysis approaches is insufficient in several domains. A thorough analysis that outlined the difficulties facing Sentiment Analysis and potential solutions for each issue. Advanced features can now be extracted from texts using a variety of methods, including Linguistic Enquiry and Word Count (LIWC). But the majority of these tools need some familiarity with programming. In this article, multi-class theory is used to identify tweets and a dictionary of knowledge and theory is used to determine the polarity of tweets. One essential algorithm in text mining is topic modelling. A theme model is a metaphor used to identify important themes in the data. Treat the text as a collection of topics in a topic structure where each topic relates to the message, is the fundamental principle. Each subject is perceived as a collection of words when utilising a topic model as a text-mining technique, and each document can be perceived as a collection of topics with varying proportions based on the frequency that terms appear. In this paper, Latent Dirichlet Allocation was implemented to extract topics from textual data. A generative probabilistic model is used to find hidden topics inside a set of documents.

The main purpose of this article is to extract content using Latent Dirichlet Allocation and perform inference on the IMDb dataset using VADER. Our objective is to accurately identify movie reviews as positive, negative or neutral based on their textual content. Furthermore, we find latent topics that are present in the reviews which provides us with a better understanding of the underlying themes and concepts that are present in the dataset. Through the integration of sentiment analysis and topic modelling our approach offers detailed insights into the sentiment conveyed within each topic. In addition, we want to show that topic modelling is more effective than standard sentiment analysis with VADER alone at lowering the dimensionality of the input data. Our goal is to find appropriate topics that capture significant themes found in the evaluations by measuring the coherence score of the extracted topics using coherence score. Incorporating topic modelling to extract more insightful information from textual data and reduce dimensionality improves computational efficiency which advances sentiment analysis algorithms.

In our approach we achieved enhanced accuracy by employing topic modelling compared to traditional sentiment analysis following the reduction of input dimensions from 50k reviews to 30 topics. This integration not only improved computational efficiency but also effectively reduced the dimensions the input data demonstrating the effectiveness of topic modelling in capturing sentiments while reducing computational complexity.

II. LITERATURE SURVEY

Article "Sentiment analysis of IMDb movie reviews using short-term neural network" [1] Sentiment analysis of IMDb movie reviews using short-term temporal (LSTM) neural network. It investigates how well long-term memory models capture the temporal dependencies in movie reviews for sentiment analysis. The study, “Movie review analysis Emotion analysis of IMDb movie reviews” applies emotion analysis methods to the analysis of IMDb movie reviews classifying the emotions represented in the Recommend using algorithms such as Naive Bayes and Support Vector Machines [2]. Taken together these two publications [3,4] offer a thorough overview of sentiment analysis and opinion mining research including key ideas techniques and developments in the area. For sentiment analysis and opinion mining tasks they investigate a variety of algorithms such as machine learning lexicon-based approaches and deep learning techniques. The authors of the paper “The Psychological Meaning of Words LIWC and Computerized Text Analysis Methods.” [5] introduced LIWC (Linguistic Enquiry and Word Count) and computerised text analysis techniques to present an innovative viewpoint on language and social psychology by revealing the psychological significance of words. Latent Dirichlet Allocation is used in “Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model” and “Sentiment Analysis and Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal for sentiment analysis and topic modelling”. This is used for tasks like topic modelling and sentiment analysis because of its capacity to reveal latent topics within textual data and its efficacy in capturing the underlying structure of the documents. [6,7]. In order to extract sentiment and topics from text data simultaneously the research “Joint Sentiment/Topic Model for Sentiment Analysis” [8] presents a novel combined sentiment and topic model that combines sentiment analysis with topic modelling. The paper “Method of Text Summarization Using Sentence Based Topic Modelling with Bert” provides a novel approach to text summarising utilising Latent Semantic Analysis (LSA) and sentence-based topic modelling with BERT. This method addresses common summary difficulties such information redundancy and content coherence [9]. “Sentiment Analysis and Topic Modelling to the Flood Disaster in Jakarta on Twitter.” In order to provide insights into public mood and theme patterns surrounding the natural calamity the study [10] “Sentiment Analysis and Topic Modelling Using the same Method related to the Flood Disaster in Jakarta on Twitter.” highlights the application of sentiment analysis and topic modelling to analyse Twitter data connected to the Jakarta flood disaster. With the goal to uncover topics and sentiments expressed in large-scale Twitter data the paper “Sentiment Analysis and Topic Modelling for Identification of Government Service Satisfaction” utilises topic modelling and sentiment analysis on global warming tweets.[11] These studies use Latent Dirichlet Allocation for sentiment analysis and topic modelling; Akhmedov et al. (2021) offer a Topic/Document/Sentence (TDS) model, and Lim and Buntine (2014) use sentiment lexicons and hashtags to extract views on products from Twitter data. They provide insights into the preferences and opinions of the general population and aid in the comprehension of sentiment and topic trends in social media debate. [12,14] The paper “Tracking mosquito-borne diseases via social media a machine learning approach to topic modelling and sentiment analysis” introduce new methods for analysing social media data use machine learning techniques for topic modelling and sentiment analysis to track mosquito-borne diseases via social media providing insights into public perceptions and trends related to public health issues use these methods to track the evolution of COVID-19 discourse in South-east Asian countries.[15,16]. This research “We-Media Facilitating Cultural Transmission: Topic Modelling and Sentiment Analysis” [17] investigates the function of We-Media in promoting cultural transmission through sentiment analysis and topic modelling. In order to analyse the sentiment of COVID-19 tweets from particular hashtags in Nigeria, this study “Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using Text Blob analyser” uses the VADER and Text Blob analysers. This illustrates the creative use of several sentiment analysis algorithms in a particular geographic setting.[18] “Public Sentiment Analysis about Independent Curriculum with Annotations using the Naive Bayes and K-Nearest Neighbour Methods” [19] paper uses annotations to analyse public opinion towards the independent curriculum. They do this by utilising the Naive Bayes and K-Nearest Neighbour techniques. Without specifically mentioning algorithms the work Visually interpret sentiment analysis results of social media posts Emphasize the use of visualization to clarify the perspective of analysing the results produced by social media posts. [20]

III. DATASET

In this paper, the IMBD movie reviews dataset is used. The 50k reviews in the dataset are divided equally between training and testing sets. There are 12.5k favourable and 12.5k unfavourable reviews in each subset. Reviews rated less than 4 stars on IMDb are classified as negative whereas those rated greater than 7 stars are deemed positive. This system is used to assign sentiment labels to reviews. This information does not include reviews with ratings other than. Furthermore, the information is organised so that each movie has a maximum of 30 reviews. The standard deviation of 172.91 words and the average review length of 234.76 words show the variation in review lengths. In addition, the dataset includes 88.5k unique words in its vocabulary demonstrating the variety of languages used in the evaluations. . Table I shows a preview of the dataset showing positive and negative reviews.

TABLE I. PREVIEW OF IMDB DATASET

Review	***Sentiment***
“Probably my all-time favourite movie, a story of selflessness, sacrifice and dedication to a noble cause, but it's not preachy or boring. It just never gets old, despite my having seen it some 15 or more times in the last 25 years.”	Positive
“It is not only oppressively preachy, but absurd, stage bound, dramatically straight-jacketed, and painfully overwrought.”	Negative

IV. METHODOLOGY

We have developed a comprehensive approach to sentiment analysis in order to enhance traditional techniques. Our methodology utilizes topic modeling to capture thematic context and nuanced sentiment patterns in textual data. This allows for more informed decision-making and insights across various domains.

A. Data Preprocessing

The 50K IMDB reviews were prepared for analysis through data preprocessing. This involved tokenizing each review into individual words and removing common, non-meaningful words (stop words) such as "the" and "is" to focus on the sentiment-carrying content. Further, words are lemmatized to their base form (e.g., "walked" becoming "walk") for improved analysis and efficiency.

B. Sentiment Analysis with LDA Topic Modelling.

Pre-processed text data from the Data Frame is used for topic modelling. By calculating correlation scores for different numbers of topics, the goal is to determine the best number of topics for a particular organization. To do this, each topic's top words for example, the top 20 words are taken into account. The pairwise similarity between these top words is used to get the coherence score. The cv coherence metric was used to assess the coherence score which is a measure of topic interpretability. The ideal number of topics for The model is determined by considering the topics that give the best results. The below table (Table II) offers a concise summary of the coherence scores attained for varying topic counts.

TABLE II. COHERENCE SCORE

Number of Topics	***Coherence Score***
5	0.2885
10	0.3105
15	0.3208
20	0.3347
25	0.3324
30	0.3238
35	0.3261

Opinion analysis on topics created with the Latent Dirichlet Allocation (LDA) model using VADER. With careful consideration to both sentiment strength and polarity VADER is specifically engineered to analyze sentiments expressed in text. Based on the top 20 terms connected to each of the 30 topics that the model retrieved sentiment scores were determined for each topic. The overall emotional tone that the topic's words convey is reflected in these sentiment scores. The emotional context of each subject is revealed by these sentiment ratings which facilitate the interpretation and comprehension of the topics produced by the model. The graph below displays the sentiment scores for each topic. Every topic in this graph is indicated by its index on the x-axis. On the y-axis the sentiment scores are shown. This chart shows the emotional scores of four different groups: neutral, negative, positive and mixed. Positive ratings indicate positive sentiment negative scores indicate negative sentiment and neutral values indicate no emotions. These scores offer insights into the general emotional tone of each topic. Sentiment scores obtained for 30 topics is depicted in the figure (Fig.1) below:

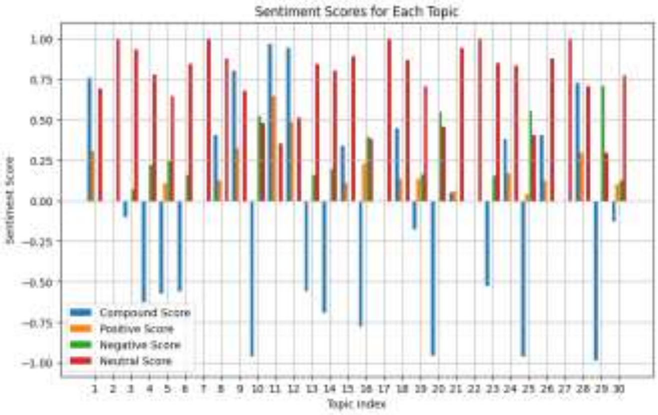

Fig.1 Sentiment Scores for Each Topic

In this paper, based on the distribution of subjects within each review the model was trained on the preprocessed review data and each review was given a dominating topic. The mapping of reviews to their corresponding prevalent topics was made easier by this method. Designating Dominant Topics-The model determined the dominant topic for every review by analyzing the topic distribution inside that review. For that review the dominant topic was determined by looking at the topic having the highest probability in the topic distribution. Sentiment Analysis for Predominant Topic- It is optional to conduct sentiment analysis on the reviews according to their predominant topic. This gives information about the sentimental tone of reviews about particular subjects. Review Mapping to Topics-Lastly each review was mapped to a prominent subject. Understanding the primary themes or topics covered in the evaluations was made easier by this mapping.

C. Determining the optimal threshlod value:

Improving the performance of emotional analysis models, we looked into the effect of different threshold values on accuracy. To achieve this, we create a loop that repeats initializing an array from 0 to 1.. For each threshold value, we calculated the sentiment analysis model's accuracy using the sentiment labels derived from the sentiment scores. The sentiment labels were created by applying a threshold to the sentiment scores obtained during the sentiment analysis process. After completing the loop, we examined the accuracy results for each threshold value. The sentiment labels were created by applying a threshold to the sentiment scores obtained during the sentiment analysis process. After completing the loop, we examined the accuracy results for each threshold value. Our findings revealed that at a threshold value of 0.83, the sentiment analysis model produced the highest accuracy as depicted below(Fig.2):

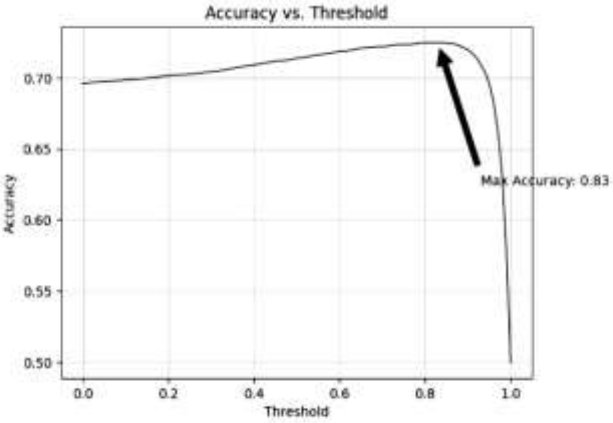

Fig.2 Accuracy vs Threshold

D. Sentiment Analysis with Vader analyser:

The textual data is preprocessed in the Data Frame, which includes handling missing values, tokenizing text by splitting it into words, removing stop words, and optional stemming or lemmatization to normalize the text. Following preprocessing, we use the Sentiment Strength Analyzer in the VADER library to calculate a sentiment score for each comment in the database. The sentiment scores, including compound scores that represent the overall sentiment polarity, are then extracted and saved in a new column called 'sentiment-compound'. Descriptive statistics for sentiment compound scores are also printed to provide an overview of the dataset's overall sentiment distribution. This method establishes a baseline for sentiment analysis, allowing for the assessment of sentiment polarity without regard for topic-based context. The below table shows the overall sentiment score without topic modelling.

V. Results and Discussion

Our analysis reveals distinct themes captured by each of the 30 topics generated using topic modelling. Each topic is associated with a group of words and their consequences, which provide insight into the dataset's thematic diversity. The summary of the top words and their probabilities for the first three identified topics. is listed in the table below (Table III)

TABLE III. Topic-wise Summary

Topic 1		***Topic 2***		***Topic 3***	
word	***score***	***word***	***score***	***word***	***score***
comedy	0.059	family	0.014	horror	0.029
funny	0.047	father	0.013	effects	0.009
game	0.019	school	0.013	gore	0.007
humor	0.015	mother	0.012	dead	0.007

The statistical summary of the overall sentiment scores, calculated without sentiment analysis, provides useful information about the sentiment distribution throughout the dataset. With 50,000 sentiment scores analyzed, the average sentiment score of 0.369 indicates a preference for positive sentiment. However, the standard deviation of 0.766 indicates significant variation in sentiment polarity across the dataset. The sentiment scores range from -0.999 to 0.999, representing both negative and positive sentiments as depicted below. (Table IV)

TABLE IV. Overall Sentiment Score without Topic Modelling

Count	*50000.0*
Mean	*0.369*
Std	*0.766*
Min	*-0.999*
Max	*0.999*

The experimental results show that including topic modeling in Sensitivity analysis improved sentiment classification accuracy, precision, recall, and F1 scores compared to standard sentiment analysis. without topic modeling. By using the semantics of data files, conceptual thinking with the model can provide a better understanding of the ideas expressed in words., resulting in more accurate sentiment predictions. The below table (Table V) presents the performance metrics for sentiment analysis with and without topic modeling, expressed as percentages.

TABLE V. Evaluation Metric

	With Topic Modelling in	***Without Topic Modelling in***

	percentage	***percentage***
Accuracy	72.46	71.17
Precision	72.73	71.4
Recall	72.7	72.25
F1 Score	72.53	71.7

VI. Conclusion and Future Works

In this paper, we looked at the effectiveness of sentiment analysis techniques with and without topic modelling using Latent Dirichlet Allocation (LDA) to extract latent topics from textual data. To sum up, our investigation into sentiment analysis through topic modelling with LDA has produced encouraging outcomes outperforming the precision attained with conventional VADER sentiment analysis on the IMDb reviews dataset. Our method performs better even though we only use 30 topics to represent the large corpus of 50k reviews. Through the incorporation of topic modelling into the sentiment analysis workflow we have optimised computational efficiency while simultaneously improving accuracy. These results highlight the effectiveness of topic modelling approaches in identifying underlying themes in textual data and capturing complex sentiment expressions. Our research suggests including topic modelling techniques into sentiment analysis to optimise precision and effectiveness while interpreting intricate textual information. Furthermore, our method dramatically decreased the input data's dimensions while retaining a high level of accuracy, demonstrating how topic modelling can improve sentiment analysis jobs. Our sentiment analysis model improved in accuracy, precision, recall and F1-score after incorporating contextual information provided by latent topics.

In order to improve the accuracy of our topic modelling approach we plan to investigate the use of more advanced topic modelling approaches like Latent Semantic Analysis (LSA) or Hierarchical Dirichlet Processes (HDP). We hope to refine the topics collected and enhance the sentiment analysis system's overall performance by utilising these advanced techniques to capture more complex semantic linkages within the textual material. We hope that this effort will improve the precision and accuracy of topic identification which will result in more intelligent analyses of the sentiment indicated in the reviews. Furthermore, examining the possible intersections between these complex topic modelling approaches and sentiment analysis techniques may provide novel insights on how to improve our understanding of the dynamics of sentiment in difficult text datasets.

References

[1] Qaisar, Saeed Mian. "Sentiment analysis of IMDb movie reviews using long short-term memory." *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*. IEEE, 2020.

[2] Topal, Kamil, and Gultekin Ozsoyoglu. "Movie review analysis: Emotion analysis of IMDb movie reviews." *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016.

[3] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and Trends® in information retrieval* 2.1–2 (2008): 1-135.

[4] Liu, Bing. *Sentiment analysis and opinion mining*. Springer Nature, 2022.

[5] Tausczik, Yla R., and James W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." *Journal of language and social psychology* 29.1 (2010): 24-54.

[6] Yang, Sidi, and Haiyi Zhang. "Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis." *International Journal of Computer and Information Engineering* 12.7 (2018): 525-529.

[7] Buhin Pandur, Maja, et al. "Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal." *SiKDD Conference on Data Mining and Data Warehouses*. Slovenian KDD Conference on Data Mining and Data Warehouses, 2021.

[8] Lin, Chenghua, and Yulan He. "Joint sentiment/topic model for sentiment analysis." *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009.

[9] Gupta, Hritvik, and Mayank Patel. "Method of text summarization using LSA and sentence based topic modelling with Bert." *2021 international conference on artificial intelligence and smart systems (ICAIS)*. IEEE, 2021.

[10] Rahmadan, M. Choirul, et al. "Sentiment analysis and topic modelling using the lda method related to the flood disaster in jakarta on twitter." *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*. IEEE, 2020

[11] Qiao, Fang, and Jago Williams. "Topic modelling and sentiment analysis of global warming tweets: Evidence from big data analysis." *Journal of Organizational and End User Computing (JOEUC)* 34.3 (2022): 1-18.

[12] Farkhod, Akhmedov, et al. "LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model." *Applied Sciences* 11.23 (2021): 11091.

[13] Yang, S., & 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.

[14] Lim, K. W., & Buntine, W. (2014, November). Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management* (pp. 1319-1328).

[15] Mathayomchan, Boonyanit, Viriya Taecharungroj, and Walanchalee Wattanacharoensil. "Evolution of COVID-19 tweets about Southeast Asian Countries: topic modelling and sentiment analyses." *Place Branding and Public Diplomacy* 19.3 (2023): 317-334.

[16] Ong, Song-Quan, and Hamdan Ahmad. "Tracking mosquito-borne diseases via social media: a machine learning approach to topic modelling and sentiment analysis." *PeerJ* 12 (2024): e17045.

[17] Zhong, Zilong. "We-Media Facilitating Cultural Transmission: LDA-based Topic Modelling and Sentiment Analysis." *London Journal of Research In Humanities and Social Sciences* 23.18 (2023): 51-64.

[18] Abiola, O., Abayomi-Alli, A., Tale, O. A., Misra, S., & Abayomi-Alli, O. (2023). Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob

[19] analyser. *Journal of Electrical Systems and Information Technology*, *10*(1), 5.

[20] Sitorus, Fernandus Paian, Ema Utami, and Mei Parwanto Kurniawan. "Public Sentiment Analysis about Independent Curriculum with VADER Annotations using the Naive Bayes and K-Nearest Neighbor Methods."

[21] Jain, Rachna, et al. "Explaining sentiment analysis results on social media texts through visualization." *Multimedia Tools and Applications* 82.15 (2023): 22613-22629.

[22] P. Kumar R, B. Mohan G, and G. D. Sai, “Ensemble Machine Learning Models in Predicting Personality Traits and Insights using Myers-Briggs Dataset,” in 2023 International Conference on Advances in Computing, Communication and Applied Informatics

(ACCAI), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ACCAI58221.2023.10199294.

[23] C. Spandana et al., “An Efficient Genetic Algorithm based Auto ML Approach for Classification and Regression,” in 2023 International Con ference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), IEEE, 2023.

[24] C. Spandana et al., “Underwater Image Enhancement and Restoration Using Cycle GAN,” in A.E. Hassanien, O. Castillo, S. Anand, A. Jaiswal (eds) International Conference on Innovative Computing and Communications. ICICC 2023. Lecture Notes in Networks and Systems, vol. 537, Springer, Singapore, 2023. doi: 10.1007/978-981-99-3010-4 9.