2024

A Comparative Analysis of the Interpretability of LDA and LLM for Topic Modeling: The Case of Healthcare Apps

Omar El-Gayar

Mohammad A. Al-Ramahi

Abdullah Wahbeh

Tareq Nasralah

Ahmed El Noshokaty

August 2024

A Comparative Analysis of the Interpretability of LDA and LLM for Topic Modeling: The Case of Healthcare Apps

Omar El-Gayar
Dakota State University, omar.el-gayar@dsu.edu

Mohammad Al-Ramahi
Texas A&M University - San Antonio, mohammad.abdel@tamusa.edu

Abdullah Wahbeh
Slippery Rock University of Pennsylvania, abdullah.wahbeh@sru.edu

Tareq Nasralah
Northeastern University, t.nasralah@northeastern.edu

Ahmed Elnoshokaty
California State University, San Bernardino, ahmed.elnoshokaty@csusb.edu

A Comparative Analysis of the Interpretability of LDA and LLM for Topic Modeling

Completed Research Full Paper

Omar El-Gayar
Dakota State University
omar.el-gayar@dsu.edu

Mohammad Al-Ramahi
Texas A&M University-San Antonio
mohammad.abdel@tamusa.edu

Abdullah Wahbeh
Slippery Rock University
abdullah.wahbeh@sru.edu

Tareq Nasralah
Northeastern University
t.nasralah@northeastern.edu

Ahmed Elnoshokaty
California State University, San Bernardino
ahmed.elnoshokaty@csusb.edu

Abstract

This study compares the interpretability of the topics resulting from three topic modeling techniques, namely, LDA, BERTopic, and RoBERTa. Using a case study of three healthcare apps (MyChart, Replika, and Teladoc), we collected 39,999, 52,255, and 27,462 reviews from each app, respectively. Topics were generated for each app using the three topic models and labels were assigned to the resulting topics. Comparative qualitative analysis showed that BERTopic, RoBERTa, and LDA have relatively similar performance in terms of the final list of resulting topics concerning human interpretability. The LDA topic model achieved the highest rate of assigning labels to topics, but the labeling process was very challenging compared to BERTopic and RoBERTa, where the process was much easier and faster given the fewer numbers of focused words in each topic. BERTopic and RoBERTa generated more cohesive topics compared to the topics generated by LDA.

Keywords

Topic modeling, Large Language Models (LLM), Latent Dirichlet Allocation (LDA).

Introduction

Topic modeling is a popular text analysis technique that provides an automatic approach for coding a collection of documents into a set of meaningful coding categories, namely topics (Mohr and Bogdanov 2013), where each topic represents a cluster of documents and words with similar meanings. Topic modeling has been widely used in the fields of natural language processing (NLP), information retrieval, and data mining and analysis to help extract insights from a collection of documents (Wang et al. 2010). Topic modeling is considered an unsupervised machine learning technique that is mainly used to uncover hidden patterns and structures from a large set of unstructured data (George and Sumathy 2023). It can organize, search, and summarize unstructured data, making it a valuable tool in various fields (Egger and Yu 2022).

Different techniques and approaches have been used to extract topics using topic modeling including Probabilistic Latent Semantic Analysis (PLSA) (George and Birla 2018; Lu et al. 2011), Latent Dirichlet Allocation (LDA) (Khadija and Nurharjadmo 2024; Lu et al. 2011), Non-negative Matrix Factorization (NMF) (Khadija and Nurharjadmo 2024; Shi et al. 2018), and BERTopic (Bidirectional Encoder Representations from Transformers) (Grootendorst 2022). LDA became a widely used algorithm for extracting meaningful topics from a collection of documents (Jelodar et al. 2019), social media data, and other types of text data (Abuzayed and Al-Khalifa 2021).

LDA aims to discover the topics in a document based on the words it contains. It is a probabilistic generative model that assigns a distribution to each document's latent topics and word allocations (Blei et al. 2003; Khadija and Nurharjadmo 2024). The probabilistic model aims at discovering the latent semantic structures or topics within a collection of unstructured text (Blei et al. 2003), which in turn helps to understand the main themes in the text collections.

A recent popular method for extracting topics from unstructured text is BERTopic. BERTopic uses BERT embeddings and class-based TF-IDF to generate dense clusters of documents and extract topic representations (Grootendorst 2022; Vahidnia et al. 2021). BERTopic is very well suited for variations of topic modeling, such as guided topic modeling and dynamic topic modeling (Wang et al. 2023). BERTopic provides good usability performance across various tasks by separating the process of clustering documents and generating topic representations (Grootendorst 2022; Hidayat et al. 2022).

A robustly optimized BERT pretraining approach proposed by Facebook is the Robustly Optimized BERT Pre-training Approach (RoBERTa), which is meant to optimize the training of BERT architecture during pre-training for modeling and learning contextual information from text (Angin et al. 2022; Li et al. 2023), question and answer system development (Suwarningsih et al. 2022), and predicting and classifying reports (Angin et al. 2022; Putra and Setiawan 2022). RoBERTa as a pre-trained model is used to learn the dynamic meaning of words in a specific context and improve the semantic representation of words (Sun and Hou 2022). It utilizes pretraining models and techniques like hyperparameter tuning, Synthetic Minority Oversampling Technique (SMOTE), and Linguistic Inquiry Word Count (LIWC) to enhance its performance (Malik et al. 2023). RoBERTa has been shown to perform well in different problem domains such as token replacement classification, personality prediction, and text classification (Gao et al. 2022).

Several studies have compared different approaches for the topic mining task (Abuzayed and Al-Khalifa 2021; Atagün et al. 2021; Axelborn and Berggren 2023; George and Sumathy 2023; Murfi et al. 2024; Prakash et al. 2023). However, these studies predominantly relied on quantitative measures with limited attention to the human interpretability of the topics generated using different topic modeling techniques. Accordingly, this study aims to evaluate the interpretability of topics generated using LDA representing a class of probabilistic topic models, and BERTopic and RoBERTa representing large language models (LLM). The evaluation leverages three different datasets representing users' reviews of three medical apps (Replika, Teladoc, MyChart) obtained from the Google Play store.

Related Work

George & Sumathy, (2023) compared a hybrid approach that utilizes BERT and LDA with traditional BERT and LDA. The authors used principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) to reduce data dimensionality. The results were evaluated using the Silhouette Score. Experimental results showed that the proposed hybrid approach utilizing clustering and dimensionality reduction could help generate more coherent topics and therefore could be used for developing topic modeling applications. Abuzayed & Al-Khalifa, (2021) have compared BERTopic with other known topic modeling techniques such as LDA, NMF. The authors used the results from LDA and NMF as a baseline and then used BERTopic with different word embedding techniques with a total number of topics ranging between 5 and 500 topics. Normalized pointwise mutual information (NPMI) was used to evaluate the results from different topic models. Experimentation and results showed that the overall topics generated by BERTopic are better than the ones obtained by LDA and NMF.

Murfi et al., (2024) extended the Eigenspace-based Fuzzy C-Means (EFCM) model using BERT for text representation in topic modeling. The authors used two variations of coherence score, the contextualized topic coherence (CTC) and the topic coherence-word2vec (TC-W2V). In addition, sensitivity analysis was performed for comparative analysis between the proposed BERT-EFCM and traditional TF-IDF. Experimental results showed that the proposed extended BERT-EFCM improves the coherence scores, especially using CTC, for topic detection compared to the traditional TF-IDF approach. Axelborn & Berggren, (2023) compared the performance of LDA and BERTopic for analyzing and categorizing textual data. Data was preprocessed and topics were obtained using LDA and BERTopic. The resulting topics were then evaluated using the perplexity and coherence scores, then the interpretability of the topics. Results showed that the quantitative analysis demonstrated that BERTopic is slightly better than LDA. However, LDA was

better than BERTopic in terms of topics quality, interpretability, and capturing meaningful and coherent topics.

Prakash et al., (2023) proposed "PromptMTopic, a novel multimodal prompt-based model designed to learn topics from both text and visual modalities" using large language models and compared the performance of the proposed models. The authors evaluated the proposed model using three real-world meme datasets and compared its performance against BERTopic, LDA, NMF, and CTM. PromptMTopic mainly utilized ChatGPT for identifying the high-level topics as well as the representative keywords for the topic. Results showed that PromptMTopic can identify relevant topics that are relevant compared to existing models. Atagün et al., (2021) compared the performance of variations of LDA and BERT for clustering tasks. Following data preprocessing and model generation, clusters were identified using the chosen models. The performance of the models was compared using the Silhouette measure. Results showed that the combined BERT, LDA, and Clustering model achieved the best performance compared to other models such as LDA with clustering, and BERT with clustering.

According to the literature, LDA and BERTopic share common characteristics, but also vary concerning the approaches and capabilities (Egger and Yu 2022; George and Sumathy 2023; Grootendorst 2022). Overall, LDA and BERTopic have their strengths and weaknesses, and the choice between them depends on the specific requirements of the topic modeling task. Furthermore, it is very challenging to find the optimal quantitative measure for evaluating and comparing the quality of topics generated using different models such as LDA and BERTopic (Abuzayed and Al-Khalifa 2021). Furthermore, the majority of the comparative studies in the literature utilized quantitative measures for comparing the performance of topic models without the qualitative analysis that involved the generation of the actual topic (Abuzayed and Al-Khalifa 2021; Atagün et al. 2021, 2021; Caliskan et al. 2022; George and Sumathy 2023; Murfi et al. 2024; Pachlore and Chakkarwar 2023). We were able to find one study that evaluated the resulting topics for interpretability (Axelborn and Berggren 2023).

Research Method

Figure 1 shows the research method for generating topics using LDA, BERTopic, and RoBERTa. The data used in this study was collected from three popular health apps on the Google Play app store. More specifically, we collected a total of 39,999 reviews about the MyChart app, an app that helps patients access health information and manage self-care. We also collected a total of 52,255 reviews about the Replika app, a popular AI companion that supports those who need mental health and emotional support. Finally, we collected a total of 27,462 reviews about the Teladoc app, a popular app that helps patients connect with service providers and receive complete care in a convenient manner. A Python script is used to preprocess the data, generate the topics, and visualize the results.

Data preprocessing is a critical step to perform topic modeling. In this study, each user review was converted to lower case. Next, text was cleaned from stop words, numbers, special characters, symbols, hashtags, mentions, and any word that is less than three characters in length. Finally, each review was lemmatized and represented as bigrams (Bekkerman and Allan 2003).

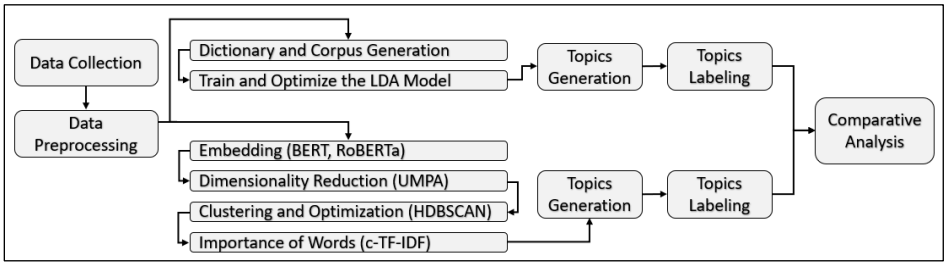


Figure 1. Research Method for Topic Generation Using LDA, BERTopic, and RoBERTa

For topic modeling using LDA, the processed reviews were used to create a dictionary, also known as a vocabulary, where each word in the dictionary is unique and assigned an index. The dictionary acts as a map where each unique term is given a specific identifier (Barde and Bainwad 2017). For each processed review, represented as bigrams, a tuple is created containing the bigram's identifier from the dictionary and the corresponding frequency in the review.

Next, the dictionary and processed reviews are used as inputs for the LDA model. For the LDA model, the number of topics needs to be specified. To do so, we optimized the topic model using the coherence score measure. "*Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic*" (Stevens et al. 2012). The coherence score was selected because it is the best measure for applications that require end-users' interaction with the generated topics (Stevens et al. 2012) and it helps provide better human interpretability of the generated topics (Röder et al. 2015) compared to other measures. Once the optimal number of topics was obtained, the LDA then uses statistical inference to assign each document a mixture of topics, and each topic a distribution of words.

For Topic modeling using BERTopic and RoBERTa, the process is similar with slight changes and improvement to the embedding step for RoBERTa. First, we need to represent each user review as a set of numerical values using BERT-base embedding and a slightly modified BERT-base embedding, for key hyperparameters and tiny embedding tweaks, for RoBERTa (McCarley et al. 2021). BERT embedding provides better representation and performance compared to traditional techniques as well as similar techniques for word embeddings (Alsentzer et al. 2019; Karande et al. 2021), especially when working with data in large sparse matrices in NLP (Karande et al. 2021).

For the next step, there is a need to reduce the dimensionality of the data from the embedding step. To do so, the UMAP (Becht et al. 2019; McInnes et al. 2018, 2020) has been used. UMPA is considered a viable technique for dimensionality reduction (McInnes et al. 2018, 2020) and is known for its speed and ability to preserve the structure of the data in lower dimensions (George and Sumathy 2023). Furthermore, topic modeling based on UMAP demonstrates better performance and helps achieve precise context-based features (George and Sumathy 2023) compared to other techniques.

The dimensionality-reduced embeddings need to be clustered using a clustering algorithm. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is used to cluster reviews with similar content, which in turn could help detect topics within the dataset (Becht et al. 2019). It performs "DBSCAN over varying epsilon values and integrates the results to find the clustering that gives the best stability over epsilon" (Malzer and Baum 2020; McInnes et al. 2017). To find the optimal number of clusters for HDBSCAN, we used the Silhouette score (Rousseeuw 1987), which can help determine the minimum cluster size , which in turn "controls the balance between the preservation of global and local structures in the low dimensional embedding" (Silveira et al. 2021).

Next, we used a variation of term frequency (TF) and inverse document frequency (IDF) (Joachims 1996) called class-based TF-IDF (c-TF-IDF) (Mazzei and Ramjattan 2022; Orellana and Bisgin 2023). To have the single representation of all reviews in a single cluster, we used the following formula:

$$c - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{r}{\sum_j^n t_i} \quad \text{..... (1)}$$

Where the word frequency, t, is extracted for each cluster, i, and then divided by the total number of words, w. Next, the total number of reviews, r, is divided by the total frequency of the word, t, across all classes, n. Finally, we generated the topics from BERTopic and RoBERTa using the top 20 keywords per topic based on the corresponding words' score in the resulting c-TF-IDF, where words appearing at the top of the list carry more weight and provide a stronger representation of the topic (Mazzei and Ramjattan 2022; Orellana and Bisgin 2023).

Once topics were obtained for LDA, BERTopic, and RoBERTa, two researchers independently labeled the resulting topics to maintain a consistent and reliable process while labeling the generated topics. Inter-rater reliability (kappa statistic) (Landis and Koch 1977) was used to evaluate the labeling process to make sure that the researcher would eventually obtain similar results. The final list of topics was labeled and merged into a higher-level topic that represents factors that affect users' acceptability and usability of mobile telehealth apps. Finally, we compared the resulting topics and compared the findings.

Results

We collected a total of 39,999 reviews for the MyChart app, 52,255 reviews for the Replika App, and 27,462 reviews for the Teladoc app. Each app reviews were analyzed using three different topic modeling approaches, namely, LDA, BERTopic, and RoBERTa. BERTopic and RoBERTa models were optimized using the Silhouette score to determine the minimum cluster size. Figures 2 and 3 show the optimization results

for different sizes of minimum cluster size. For MyChart reviews, the best minimum cluster sizes were 8 and 11 with Silhouette scores of 0.954 and 0.966 for BERTopic and RoBERTa, respectively. For Replika reviews, the best minimum cluster sizes were 11 and 12, with Silhouette scores of 0.947 and 0.958, for BERTopic and RoBERTa respectively. Finally, for Teladoc reviews, the best minimum cluster sizes were 25 and 8 with Silhouette scores of 0.936 and 0.956, for BERTopic and RoBERTa respectively.

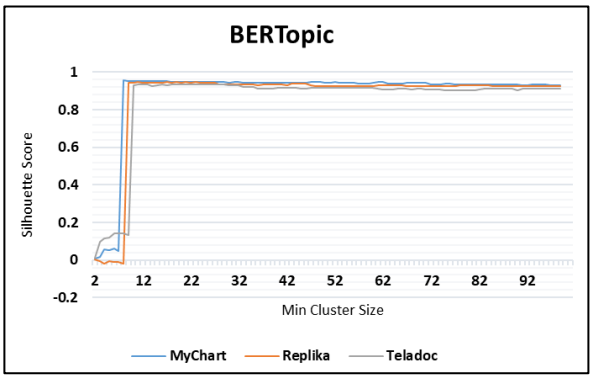


Figure 2. Silhouette score for the optimal minimum cluster size

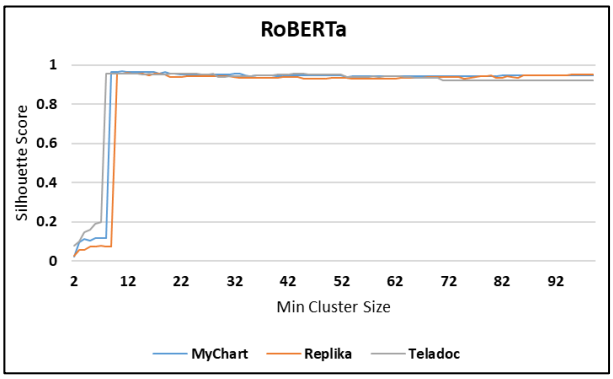


Figure 3. Silhouette Score for the optimal minimum cluster size

The LDA model was optimized using the Coherence score for determining the optimal number of topics. Figure 4 shows the results for different number of topics. For MyChart reviews using LDA, the best number of topics was 94 with a coherence score of 0.608. For Replika reviews using LDA, the best number of topics was 97 with a coherence score of 0. 628. Finally, for Teladoc reviews using LDA, the best number of topics was 81 with a coherence score of 0.612.

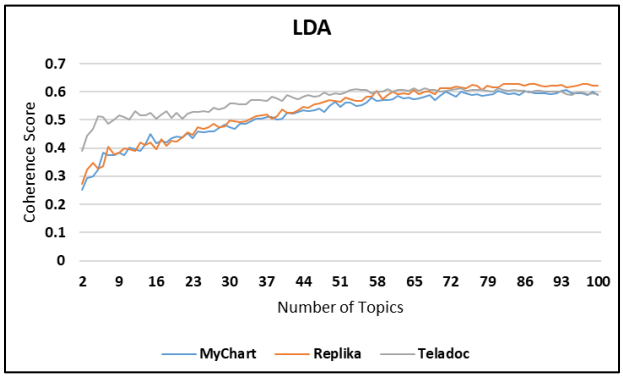


Figure 4. Coherence score for optimal number of topics

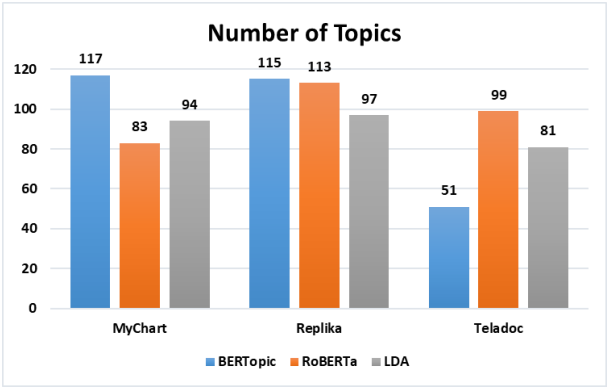


Figure 5. Number of topics for each app/model

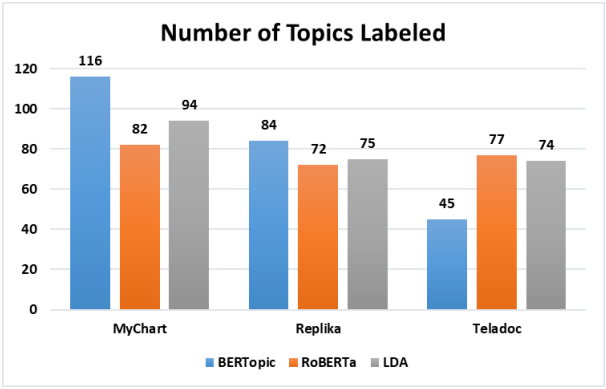


Figure 6. Number of labeled topics for each app/model

Figures 5 and 6 show the number of topics generated and number of topics labeled. As shown in the figures, for MyChart, we were able to label 116/117 (99.1%) topics using BERTopic, 82/83 (98.7) topics using RoBERTa, and 94/94 (100%) topics using LDA. For Rplika, we were able to label 84/115 (73.0%) topics using BERTopic, 72/113 (63.7) topics using RoBERTa, and 75/97 (77.3%) topics using LDA. Finally, for Teladoc, we were able to label 45/51 (88.2%) topics using BERTopic, 77/99 (77.7) topics using RoBERTa, and 74/81 (91.4%) topics using LDA.

Table 1 shows the high-level topics obtained for the Teladoc app using BERTopic, RoBERTa, and LDA. As shown in the table, we were able to obtain the same number of high-level topics as well as the same topic using the three models. The two independent researchers labeling the topics achieved a kappa statistic of 91.5% which reflect perfect agreement among different raters (Landis and Koch 1977).

BERTopic	RoBERTa	LDA
Convenience and Efficiency	Convenience and Efficiency	Convenience and Efficiency
Useful and Helpful	Useful and Helpful	Useful and Helpful
Easy to Use	Easy to Use	Easy to Use

Table 1. Teladoc high level topics obtained from each model.

Table 2 shows the high-level topics obtained for the Replika app using BERTopic, RoBERTa, and LDA. As shown in the table, we were able to obtain the same number of high-level topics as well as the same topic using RoBERTa and LDA models. However, using the BERTopic model, we ended up with one more high-level topic, while obtaining other high-level topics like RoBERTa and LDA.

BERTopic	RoBERTa	LDA
Social Support	Social Support	Social Support
Engaging Conversation	Engaging Conversation	Engaging Conversation
Fun, Entertaining, and Interesting	Fun, Entertaining, and Interesting	Fun, Entertaining, and Interesting
Cost and Subscription	Cost and Subscription	Cost and Subscription
Intelligent Learning	----	----
Technical Issues and Problems	Technical Issues and Problems	Technical Issues and Problems
Usefulness	Usefulness	Usefulness
User Friendly	User Friendly	User Friendly

Table 2. Replika high-level topics obtained from each model.

Table 3 shows the high-level topics obtained for the MyChart app using BERTopic, RoBERTa, and LDA. As shown in the table, we were able to obtain the same number of high-level topics as well as the same topic using the three models.

BERTopic	RoBERTa	LDA
Easy to Use	Easy to Use	Easy to Use
Usefulness	Usefulness	Usefulness
Appointment Management	Appointment Management	Appointment Management
Accessibility to medical information	Accessibility to medical information	Accessibility to medical information
Customer Support	Customer Support	Customer Support
Self-Monitoring/Tracking	Self-Monitoring/Tracking	Self-Monitoring/Tracking
Informative Information	Informative Information	Informative Information
User Reported Issues	User Reported Issues	User Reported Issues
Communication with Provider	Communication with Provider	Communication with Provider

Table 3. MyChart high-level topics obtained from each model.

Table 4 shows sample word clouds for the low-level topics obtained from each model. As shown in the table, we can see that BERTopic and RoBERTa are better than LDA when it comes to topic cohesiveness. As demonstrated in the sample word clouds, the clouds for BERTopic and RoBERTa consist of a set of words that are closely related in terms of their thematic content, which in turn makes the labeling process much easier and faster than the topics obtained from LDA. Such difference between BERTopic and RoBERTa on one side and LDA on the other side makes the process of interpreting and understanding the underlying theme easier using BERTopic and RoBERTa.

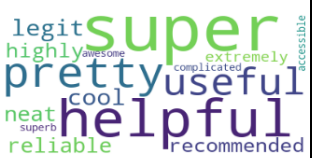
Model/Label	Ease of Use	Useful and Helpful	Cost and Subscription
BERTopic			
RoBERTa			
LDA			

Table 4. Sample Word Clouds for Low Level Topics Obtained from Each Model

Discussion

This study aims to evaluate the interpretability of topics generated using LDA representing a class of probabilistic topic models, and BERTopic and RoBERTa representing large language models (LLM). The evaluation leverages three different public opinion datasets about medical apps obtained from the Google Play store. In terms of performance, BERTopic, RoBERTa, and LDA have relatively similar performance in terms of the final list of resulting topics with respect to human interpretability. This is an interesting finding since the literature reported various interpretations concerning different topic models' performance. For example, quantitative analysis showed that clustering with dimensionality reduction in BERT could help generate more coherent topics compared to LDA and NMF (Abuzayed and Al-Khalifa 2021; George and Sumathy 2023). BERT embedding for topic modeling provided a better interpretation of patients' perspectives and perceptions that could improve the quality of care (Osváth et al. 2023).

On average, we were able to label 86.76% of the topics generated using BERTopic, 80.03% of the topics generated using RoBERTa, and 89.5% of the topics generated using LDA. These averages are comparable to the literature (Rijcken et al. 2023), where a domain expert was able to label 75% of the topics generated. The LDA topic model achieved the highest rate of labeling given the fact that such a model generates topics that consist of a large number of keywords in each topic. While this makes the labeling process achieve a high labeling rate using LDA compared to BERTopic and RoBERTa, the process itself was very challenging given the number of words in each topic, and the overlapping topic labels in the same topic as shown in Table 4. On the other hand, despite the low labeling rate for BERTopic and RoBERTa, the labeling process was much easier and faster given the few number of focused words in each topic, which facilitates the process of assigning a topic label. In general, our findings suggest that BERTopic and RoBERTa have shown better generalization, topic coherence, and diversity compared to LDA. This is in line with existing literature, especially for short texts (Zhou et al. 2022) like user reviews.

Further, the results showed that BERTopic and RoBERTa generated more cohesive topics that were more focused and easier to interpret compared to the topics generated by LDA. This is in line with existing literature where LLM-based models generate highly clustered embeddings that help generate topics with superior clusterability and improved semantic coherence when compared to traditional methods like LDA

(Xu et al. 2023). Quantitative comparative performance analysis of the LDA and LLM-based models showed that BERT representation of topics improves topic coherence and overall interpretability of the topics (Murfi et al. 2024). LLM-based topic models, such as RoBERTa and BERTopic, can recognize the nuances and subtleties in the data, which in turn could help identify coherent and meaningful topics that other models, such LDA, might overlook (Prakash et al. 2023).

Conclusion

We evaluated the interpretability of LDA, BERTopic, and RoBERTa for topic models. In general, we found that BERTopic and RoBERTa perform better than LDA with respect to topic interpretability. This is mainly related to the fact that topics based on LLM models are more coherent and simpler to interpret compared to traditional methods such as LDA and LSA. Furthermore, BERTopic and RoBERTa demonstrated competitive and stable performance compared to LDA when it comes to the interpretability of the topics as demonstrated in the three cases. Finally, LDA often lacks clear semantic information and has feature sparsity problems. On the other hand, BERTopic and RoBERTa are less sensitive to these problems. This work is not without any limitations. First, the HDBSCAN algorithm classifies many of the reviews as outliers. However, this was not an issue in the current work given the short nature of the reviews as well as the comparable resulting high-level topics from all models. Some apps resulted in fewer high-level topics compared to other apps. This could be attributed to the nature of the data or the users' reviews being focused on these aspects of the app. Future research may explore the performance of other probabilistic topic models such as LSA and PLSA compared to LLM-based topic models.

References

Abuzayed, A., and Al-Khalifa, H. 2021. "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," *Procedia Computer Science* (189), AI in Computational Linguistics, pp. 191–194. (https://doi.org/10.1016/j.procs.2021.05.096).

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. B. A. 2019. *Publicly Available Clinical BERT Embeddings*, arXiv. (https://doi.org/10.48550/arXiv.1904.03323).

Angin, M., Taşdemir, B., Yılmaz, C. A., Demiralp, G., Atay, M., Angin, P., and Dikmener, G. 2022. "A RoBERTa Approach for Automated Processing of Sustainability Reports," *Sustainability* (14:23), Multidisciplinary Digital Publishing Institute, p. 16139. (https://doi.org/10.3390/su142316139).

Atagün, E., Hartoka, B., and Albayrak, A. 2021. "Topic Modeling Using LDA and BERT Techniques: Teknofest Example," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, , September, pp. 660–664. (https://doi.org/10.1109/UBMK52708.2021.9558988).

Axelborn, H., and Berggren, J. 2023. "Topic Modeling for Customer Insights: A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls," UMEA University.

Barde, B. V., and Bainwad, A. M. 2017. "An Overview of Topic Modeling Methods and Tools," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, , June, pp. 745–750. (https://doi.org/10.1109/ICCONS.2017.8250563).

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. 2019. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP," *Nature Biotechnology* (37:1), Nature Publishing Group, pp. 38–44. (https://doi.org/10.1038/nbt.4314).

Bekkerman, R., and Allan, J. 2003. "Using Bigrams in Text Categorization," No. Technical Report IR-408, Center of Intelligent Information Retrieval: UMass Amherst, p. 10.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3:Jan), pp. 993–1022.

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., and Banaji, M. R. 2022. *Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics*, presented at the Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 156–170.

Egger, R., and Yu, J. 2022. "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology* (7), Frontiers Media SA. (https://doi.org/10.3389/fsoc.2022.886498).

Gao, L., Zhang, Lijuan, Zhang, Lei, and Huang, J. 2022. "RSVN: A RoBERTa Sentence Vector Normalization Scheme for Short Texts to Extract Semantic Information," *Applied Sciences* (12:21), Multidisciplinary Digital Publishing Institute, p. 11278. (https://doi.org/10.3390/app122111278).

George, L. E., and Birla, L. 2018. "A Study of Topic Modeling Methods," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, , June, pp. 109–113. (https://doi.org/10.1109/ICCONS.2018.8663152).

George, L., and Sumathy, P. 2023. "An Integrated Clustering and BERT Framework for Improved Topic Modeling," *International Journal of Information Technology* (15:4), pp. 2187–2195. (https://doi.org/10.1007/s41870-023-01268-w).

Grootendorst, M. 2022. *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*, arXiv. (https://doi.org/10.48550/arXiv.2203.05794).

Hidayat, A. A., Nirwantono, R., Budiarto, A., and Pardamean, B. 2022. "BERT-Based Topic Modeling Approach for Malaria Research Publication," in *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, , November, pp. 326–331. (https://doi.org/10.1109/ICIMCIS56303.2022.10017743).

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. 2019. "Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey," *Multimedia Tools and Applications* (78:11), pp. 15169–15211. (https://doi.org/10.1007/s11042-018-6894-4).

Joachims, T. 1996. "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.," Carnegie-mellon univ pittsburgh pa dept of computer science.

Karande, H., Walambe, R., Benjamin, V., Kotecha, K., and Raghu, T. 2021. "Stance Detection with BERT Embeddings for Credibility Analysis of Information on Social Media," *PeerJ Computer Science* (7), PeerJ Inc., p. e467.

Khadija, M. A., and Nurharjadmo, W. 2024. "Enhancing Indonesian Customer Complaint Analysis: LDA Topic Modelling with BERT Embeddings," *SINERGI* (28:1), Mercu Buana University, pp. 153–162. (https://doi.org/10.22441/sinergi.2024.1.015).

Landis, J. R., and Koch, G. G. 1977. "The Measurement of Observer Agreement for Categorical Data.," *Biometrics* (33:1), pp. 159–174. (https://doi.org/10.2307/2529310).

Li, M., Qin, Y., and Huangfu, W. 2023. "RoBERTa: An Efficient Dating Method of Ancient Chinese Texts," in *Chinese Lexical Semantics*, Lecture Notes in Computer Science, Q. Su, G. Xu, and X. Yang (eds.), Cham: Springer Nature Switzerland, pp. 293–301. (https://doi.org/10.1007/978-3-031-28956-9_23).

Lu, Y., Mei, Q., and Zhai, C. 2011. "Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA," *Information Retrieval* (14:2), pp. 178–203. (https://doi.org/10.1007/s10791-010-9141-9).

Malik, M. S. I., Imran, T., and Mamdouh, J. M. 2023. "How to Detect Propaganda from Social Media? Exploitation of Semantic and Fine-Tuned Language Models," *PeerJ Computer Science* (9), PeerJ Inc., p. e1248. (https://doi.org/10.7717/peerj-cs.1248).

Malzer, C., and Baum, M. 2020. "A Hybrid Approach To Hierarchical Density-Based Cluster Selection," in *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, , September, pp. 223–228. (https://doi.org/10.1109/MFI49285.2020.9235263).

Mazzei, D., and Ramjattan, R. 2022. "Machine Learning for Industry 4.0: A Systematic Review Using Deep Learning-Based Topic Modelling," *Sensors* (22:22), Multidisciplinary Digital Publishing Institute, p. 8641. (https://doi.org/10.3390/s22228641).

McCarley, J. S., Chakravarti, R., and Sil, A. 2021. *Structured Pruning of a BERT-Based Question Answering Model*, arXiv. (https://doi.org/10.48550/arXiv.1910.06360).

McInnes, L., Healy, J., and Astels, S. 2017. "Hdbscan: Hierarchical Density Based Clustering.," *J. Open Source Softw.* (2:11), p. 205.

McInnes, L., Healy, J., and Melville, J. 2020. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv. (https://doi.org/10.48550/arXiv.1802.03426).

McInnes, L., Healy, J., Saul, N., and Großberger, L. 2018. "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software* (3:29).

Mohr, J. W., and Bogdanov, P. 2013. "Introduction—Topic Models: What They Are and Why They Matter," *Poetics* (41:6), Topic Models and the Cultural Sciences, pp. 545–569. (https://doi.org/10.1016/j.poetic.2013.10.001).

Murfi, H., Agung, Y., Nurrihmah, S., Satria, Y., Zain, C., and Rahayu, D. 2024. "Eigenspace-Based Fuzzy C-Means with Large Language Model BERT for Topic Detection," , January 10. (https://doi.org/10.21203/rs.3.rs-3637575/v1).

Orellana, S., and Bisgin, H. 2023. "Using Natural Language Processing to Analyze Political Party Manifestos from New Zealand," *Information* (14:3), Multidisciplinary Digital Publishing Institute, p. 152. (https://doi.org/10.3390/info14030152).

Osváth, M., Yang, Z. G., and Kósa, K. 2023. "Analyzing Narratives of Patient Experiences: A BERT Topic Modeling Approach," *Acta Polytechnica Hungarica* (20:7), pp. 153–171. (https://doi.org/10.12700/APH.20.7.2023.7.9).

Pachlore, A., and Chakkarwar, V. 2023. *Opinion Mining and Tweet Analysis Using Topic Modeling by LDA with BERT and GLOVE Embedding*, presented at the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022), Atlantis Press, August 10, pp. 660–673. (https://doi.org/10.2991/978-94-6463-196-8_50).

Prakash, N., Wang, H., Hoang, N. K., Hee, M. S., and Lee, R. K.-W. 2023. "PromptMTopic: Unsupervised Multimodal Topic Modeling of Memes Using Large Language Models," in *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, New York, NY, USA: Association for Computing Machinery, October 27, pp. 621–631. (https://doi.org/10.1145/3581783.3613836).

Putra, R. P., and Setiawan, E. B. 2022. *RoBERTa as Semantic Approach for Big Five Personality Prediction Using Artificial Neural Network on Twitter*, presented at the 2022 International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS), IEEE, pp. 1–6.

Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., and Kaymak, U. 2023. *Towards Interpreting Topic Models with ChatGPT: The 20th World Congress of the International Fuzzy Systems Association*.

Röder, M., Both, A., and Hinneburg, A. 2015. "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, New York, NY, USA: Association for Computing Machinery, February 2, pp. 399–408. (https://doi.org/10.1145/2684822.2685324).

Rousseeuw, P. J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics* (20), pp. 53–65. (https://doi.org/10.1016/0377-0427(87)90125-7).

Shi, T., Kang, K., Choo, J., and Reddy, C. K. 2018. "Short-Text Topic Modeling via Non-Negative Matrix Factorization Enriched with Local Word-Context Correlations," in *Proceedings of the 2018 World Wide Web Conference*, WWW '18, Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, April 10, pp. 1105–1114. (https://doi.org/10.1145/3178876.3186009).

Silveira, R., Fernandes, C., Neto, J. A. M., Furtado, V., and Pimentel Filho, J. E. 2021. "Topic Modelling of Legal Documents via Legal-Bert," *Proceedings Http://Ceur-Ws Org ISSN* (1613), p. 0073.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. 2012. "Exploring Topic Coherence over Many Models and Many Topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea: Association for Computational Linguistics, July, pp. 952–961. (https://aclanthology.org/D12-1087).

Sun, J., and Hou, X. 2022. "A Roberta-Seq2Seq Based Model for Chinese Text Abstractive Summarization," in *International Conference on Advanced Algorithms and Neural Networks (AANN 2022)* (Vol. 12285), SPIE, June 15, pp. 261–266. (https://doi.org/10.1117/12.2637174).

Suwarningsih, W., Pramata, R. A., Rahadika, F. Y., and Purnomo, M. H. A. 2022. "RoBERTa: Language Modelling in Building Indonesian Question-Answering Systems," *TELKOMNIKA (Telecommunication Computing Electronics and Control)* (20:6), pp. 1248–1255.

Vahidnia, S., Abbasi, A., and Abbass, H. A. 2021. "Embedding-Based Detection and Extraction of Research Topics from Academic Documents Using Deep Clustering," *Journal of Data and Information Science* (6:3), pp. 99–122.

Wang, W., Mamaani Barnaghi, P., and Bargiela, A. 2010. "Probabilistic Topic Models for Learning Terminological Ontologies," *IEEE Transactions on Knowledge and Data Engineering* (22:7), pp. 1028–1040. (https://doi.org/10.1109/TKDE.2009.122).

Wang, Z., Chen, Jing, Chen, Jiangping, and Chen, H. 2023. "Identifying Interdisciplinary Topics and Their Evolution Based on BERTopic," *Scientometrics*, Springer, pp. 1–26.

Xu, W., Hu, W., Wu, F., and Sengamedu, S. 2023. "DeTiME: Diffusion-Enhanced Topic Modeling Using Encoder-Decoder Based LLM," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9040–9057. (https://doi.org/10.18653/v1/2023.findings-emnlp.606).

Zhou, M., Kong, Y., and Lin, J. 2022. *Financial Topic Modeling Based on the BERT-LDA Embedding*, presented at the 2022 IEEE 20th International Conference on Industrial Informatics (INDIN), IEEE, pp. 495–500.