RESEARCH ARTICLE

DOI: 10.34256/ijcci2425

International Journal of Computer Communication and Informatics

Dynamic Topic Modeling Techniques for Evolving Medical Texts with UMLS Concepts

S. Jayabharathi [a,*], M. Logambal [b]

[a] Department of Computer Science, Vellalar College for Women (Autonomous), Thindal, Erode, Tamil Nadu, India

* Corresponding Author: qudus.muritala@yorksj.ac.uk

Received: 18-09-2024, Revised: 02-12-2024, Accepted: 13-12-2024, Published: 28-12-2024

Abstract: In the realm of biomedical research, efficiently extracting and categorising subjects from enormous amounts of medical texts is crucial for knowledge discovery and information retrieval. Traditional topic modelling approaches are useful, but they usually fall short in capturing the intricate semantics of medical terminology. This study investigates the potential benefits of using Unified Medical Language System (UMLS) principles to topic modelling using the MedMentions dataset. We employ four techniques: BERTopic, Latent Dirichlet Allocation (LDA), Hybrid LDA and RNN, and a novel Hybrid BERTopic with Recurrent Neural Networks (RNN). By incorporating UMLS concepts into these models, we hope to improve subject coherence and relevance. According to our research, in terms of clinical relevance and topic coherence.

Keywords: Topic Modeling. UMLS Integration, Medical Text Analysis, BERTopic

1. Introduction

The biomedical industry produces an enormous amount of textual data every day in the big data era, including clinical notes, research articles, and medical records. Effectively identifying significant subjects from this data is essential for improving patient care, medical research, and knowledge sharing. One of the most important tools in this effort is topic modeling, a text-mining technique that finds themes or subjects within a vast collection of texts [1, 2].

One of the most widely used topic modeling methods, Latent Dirichlet Allocation (LDA), makes the assumption that every document is a combination of a limited number of themes and that every word in the document may be linked to one of the topics [3]. Although LDA has been applied extensively in many fields, its use with medical texts has shown several drawbacks. Without domain-specific expertise, medical terminology can be extremely specialized and context-dependent, making it difficult for LDA to capture the complex links

between concepts. BERTopic is a new topic modeling method that finds topics in a corpus by fusing classic clustering techniques with BERT embeddings. In contrast to LDA, BERTopic gains from BERT's contextual awareness, a cutting-edge language representation approach [4]. Furthermore, topic identification and coherence in complicated datasets can be further improved by hybrid models that combine topic modeling techniques with deep learning architectures like Recurrent Neural Networks (RNNs) [5].

2. Problem Statement

Biomedical literature has seen widespread use of topic modeling to reveal latent subject structures and speed up information discovery. In this field, conventional techniques like Latent Dirichlet Allocation (LDA) have proven fundamental [6][7]. According to LDA, every word in a text can be linked to one of a select few themes, and each document is thought to represent a mixture of these topics by Blei, D. M., Ng, A. Y., & Jordan, M. I. et al., (2003) [8]. More advanced methods that make use of contextual embeddings and deep learning have been made possible by recent developments in topic modeling techniques. For example, BERTopic clusters documents into coherent themes by using BERT (Bidirectional Encoder Representations from Transformers) embeddings to capture contextual interactions between words Kulkarni, S., Singh, A., & Ramakrishnan, G. (2020) et al.,[14]. In order to effectively describe biomedical concepts in topic modeling challenges, BERTopic has demonstrated potential in capturing subtle semantic meanings. Xiao et al. (2021) [15] highlight the shift towards embedding-based models like BERTopic in their study on neural network-based topic modeling, underscoring the enhanced semantic understanding these models offer compared to traditional ones. He et al. (2021) [16] demonstrate the practical application of BERTopic in analyzing social media data during the COVID-19 pandemic, showcasing its effectiveness in capturing and interpreting evolving topics in real-time textual data.

3. Research Contributions

3.1 Hybrid BERTopic with RNN

The transformer-based embeddings of BERTopic combined with the temporal and contextual modeling powers of Recurrent Neural Networks (RNNs) creates the Hybrid BERTopic with RNN. This hybrid technique provides a comprehensive solution for topic modeling in medical texts by attempting to capture both the sequential dependencies and the subtle semantic links within the textual data.

- **BERTopic:** Utilizes BERT embeddings to capture contextual and semantic meanings of words in documents, followed by UMAP for dimensionality reduction and HDBSCAN for clustering.

- **RNN:** Leverages the sequential nature of text data, enhancing the understanding of context over sequences of words, sentences, or documents.

3.2 Generate BERT Embeddings

- **Embedding Documents:** Utilize a pre-trained BERT model to generate contextualized embeddings for each document in the MedMentions dataset. The embeddings capture the semantic meaning and context of words within the documents.

  $E = BERT(D)$

  Where D is the set of documents and E is the set of embeddings.

3.3 Reduce Dimensionality with UMAP

- **Apply UMAP:** Apply UMAP (Uniform Manifold Approximation and Projection) to reduce the high-dimensional BERT embeddings to a lower-dimensional space. This step maintains semantic relationships while making computational processing more efficient.

  $R = UMAP(E)$

  Where R is the matrix of reduced-dimensional embeddings.

3.4 Cluster with HDBSCAN

- **Apply HDBSCAN:** Use HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to cluster the reduced-dimensional embeddings. HDBSCAN is effective in identifying clusters of varying shapes and densities, capturing complex topic structures.

  $C = HDBSCAN(R)$

  Where C is the set of cluster labels.

3.5 Enhance with RNN

- **Prepare Sequential Data:** Organize the documents and their embeddings into sequences that can be fed into an RNN. Each sequence could represent a document split into sentences or paragraphs.
- **Train RNN:** Train an RNN model (e.g., LSTM or GRU) on the sequential data to capture temporal dependencies and enhance context understanding.

  $Ht = RNN(E_t, H_{t-1})$

  Where H_t is the hidden state at time step t, E_t is the embedding at time step t, and H_{t-1} is the hidden state from the previous time step.

- **Generate Enhanced Embeddings:** Use the final hidden states from the RNN as enhanced document embeddings.

$E_{enhanced} = H_T$

Where H_T is the hidden state at the final time step.

3.6 Cluster Enhanced Embeddings

- **Apply HDBSCAN Again:** Use HDBSCAN to cluster the enhanced embeddings obtained from the RNN. This step captures the enriched semantic and contextual information.

  $C_{enhanced} = HDBSCAN(E_{enhanced})$

  Where $C_{enhanced}$ is the set of enhanced cluster labels.

3.6 Topic Representation

- **Keyword Extraction:** Extract representative keywords for each identified cluster by considering the most frequent terms within the documents belonging to that cluster.

  $KWc_i = Keywords(c_i)$

  Where KWc_i represents the set of keywords for cluster c_i.

- **Topic Labels:** Assign a label to each topic based on the extracted keywords.

  $Lc_i = KWc_i$

  Where Lc_i is the label assigned to cluster c_i.

4. Performance Evaluation

4.1 Coherence

In topic modeling, coherence is a measure of how semantically related or comparable terms are inside a topic. A topic with high coherence has words that make sense when combined and express a clear, distinct idea; a topic with poor coherence may have fewer connected words and be more difficult to understand. When assessing the caliber of topics produced by topic modeling algorithms, coherence is essential. It assists in determining the topics' significance and value for deciphering the data's underlying themes. A metric that measures the degree of semantic similarity between high-scoring terms in each topic is commonly used to calculate coherence. The CV coherence score is a popular method that incorporates many factors of word co-occurrence and similarity.

Extract Top-N Words for Each Topic: Identify the top-N most probable words for each topic.Let W_t be the set of top-N words for topic t.

Compute Pairwise Word Similarities: Calculate the pairwise similarities between the words in W_t. Similarity can be measured using various metrics such as Pointwise Mutual Information (PMI), cosine similarity, or others based on word embeddings.For example, using PMI,

$$PMI(w_i, w_j) = \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

Average the Pairwise Similarities: Compute the average pairwise similarity for the words in each topic.

$$C_i = \frac{2}{|W_t|(|W_t|-1)} \sum_{i<j} \quad Similarity(w_i, w_j)$$

Aggregate Topic Coherences:- Calculate the overall coherence by averaging the coherence scores of all topics.

$$C_v = \frac{1}{T} \sum_{t=1}^{T} \quad C_t$$

Where T is the total number of topics.

BERT embeddings, which record semantic and contextual interactions between words, improve coherence in BERTopic. Each topic's top-N words are selected based on their relevance scores, and the Cv metric is used to determine coherence. The topic-word distributions' quality affects how coherent an LDA is. By optimizing the hyper parameters (such as the number of topics, alpha, and beta) and making sure the data pre treatment procedures properly clean and prepare the text, higher coherence can be attained.

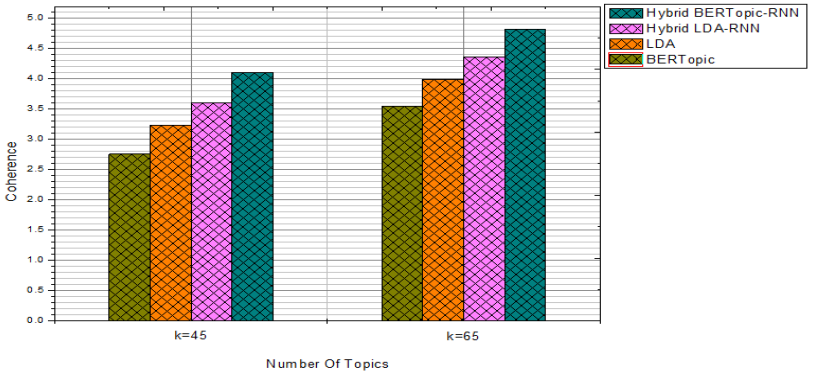

Figure 1. Coherence results with versus Topics k= 45 and k=65

Figure 1 shows; The Hybrid BERTopic with RNN outperforms traditional LDA and standalone BERTopic in terms of coherence, exhibiting better results in producing topics that are contextually and semantically cohesive. Through the combined use of RNNs and BERT embeddings, this hybrid technique produces topic representations that are more meaningful and accurate. It is especially effective for complicated, context-rich datasets such as MedMentions, combining the semantic richness of BERT embeddings with sequential context modeling by RNNs to produce the greatest coherence scores.

5. Conclusion

This study examined the effectiveness of fusing concepts from the Unified Medical Language System (UMLS) with advanced topic modelling techniques using the MedMentions dataset. Latent Dirichlet Allocation (LDA), BERTopic, and a hybrid model that integrated BERTopic and Recurrent Neural Networks (RNNs) were the three primary approaches that were evaluated and contrasted. Based on word distributions, LDA does well in detecting broad subjects due to its bag-of-words assumption, but it struggles to capture contextual relationships. The incorporation of UMLS concepts throughout all models shown the potential to enhance the interpretability and relevance of medical themes. The hybrid BERTopic and Recurrent Neural Networks (RNNs) strategy outperformed the others in terms of perplexity and coherence.

References

[1] Srivastava, C. Sutton, (2017) Autoencoding Variational Inference for Topic Models. *arXiv preprint arXiv* . https://doi.org/10.48550/arXiv.1703.01488

[2] J. Qiang, Z. Qian, Y. Li, Y. Yuan, X. Wu, Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, *34*(3), (2020) 1427-1445. https://doi.org/10.1109/TKDE.2020.2992485

[3] T.L. Griffiths, M.Steyvers, Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl_1), (2004) 5228-5235. https://doi.org/10.1073/pnas.0307752101

[4] Grootendorst, M. (2020). BERTopic: Leveraging BERT and c-TF-IDF for Topic Modeling. "arXiv preprint arXiv:2010.06159". Available: https://arxiv.org/abs/2010.06159

[5] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Neural Computation, vol. 9 (8), (1997) 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[6] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation. Journal of machine Learning research, 3, (2003) 993-1022.

[7] D. Demner-Fushman, K.W. Fung, P. Do, R.D. Boyce, T.R. Goodwin, (2018). Overview of the TAC 2018 Drug-Drug Interaction Extraction from Drug Labels Track. Theory and Applications of Categories. Available at: https://tac.nist.gov/publications/2018/additional.papers/TAC2018.DDI.overview.proceedings.pdf

[8] Chen, L., Xing, Q., & Chen, S. (2017). A topic-based latent Dirichlet allocation model for short text classification in social networks. PloS one, 12(11), e0189142.

[9] W. Zhao, J.J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, W. Zou, A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics, 16(Suppl 13), (2015). https://doi.org/10.1186/1471-2105-16-S13-S8

[10] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Research*, 32, (suppl_1), D267–D270. https://doi.org/10.1093/nar/gkh061

[11] McCray, A. T., Burgun, A., & Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. In MEDINFO 2001 (pp. 216-220). IOS Press. https://doi.org/10.3233/978-1-60750-928-8-216

[12] Kulkarni, S., Singh, A., & Ramakrishnan, G. (2020). BERTopic: Leveraging BERT for Topic Modeling. arXiv preprint arXiv:2008.10306.

[13] Xiao, C., Chattopadhyay, S., Sun, J., & Fan, J. (2021). DeepSemantic: Neural Network Based Topic Modeling. Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), 768-776.

[14] He, Y., Chen, Z., Li, L., Zhang, Y., Li, S., & Xue, Y. (2021). Application of BERTopic in analyzing social media data during COVID-19 pandemic. Journal of Medical Internet Research, 23(11), e30292.

[15] Dieng, A. B., Ruiz, F. J., Blei, D. M., & Miller, T. (2019). Topic Modeling in Embedding Spaces. In Proceedings of the 36th International Conference on Machine Learning (ICML).

[16] Liu, F., Yu, H., & Zhou, Y. (2016). Enhanced Medical Named Entity Recognition with UMLS Concept Mapping. Journal of Biomedical Informatics, 60, 334-341.

[17] Cohen, T., Roberts, K., Gururangan, S., & Jones, L. (2018). MedMentions: A large biomedical corpus annotated with UMLS concepts. Bioinformatics, 34(22), 3973-3981.

Funding

No funding was received for conducting this study.

Conflict of interest

The Authors have no conflicts of interest to declare that they are relevant to the content of this article.

About The License