A Comparative Analysis of Topic Modelling Methods for Temporal Analysis of News Articles

1st Md. Anonto Shuvo
Dept. of CSE
BRAC University
Dhaka, Bangladesh
md.anonto.shuvo@g.bracu.ac.bd
ID: 23141036

2nd Aditi Saha Ria
Dept. of CSE
BRAC University
Dhaka, Bangladesh
aditi.saha.ria@g.bracu.ac.bd
ID: 20101238

3rd Mahin Shahriar Efaz
Dept. of CSE
BRAC University
Dhaka, Bangladesh
mahin.shahriar.efaz@g.bracu.ac.bd
ID: 20101225

4thMd. Sabbir Hossain
Dept. of CSE
BRAC University
Dhaka, Bangladesh
ext.sabbir.hossain@bracu.ac.bd

5th Md. Farhadul Islam
Dept. of CSE
BRAC University
Dhaka, Bangladesh
md.farhadul.islam@g.bracu.ac.bd

6th Annajiat Alim Rasel
Dept. of CSE
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—In order to analyze and comprehend temporal patterns from large datasets, various learning models have been used. Among these models, Topic Modelling has been very useful for analyzing latent topics from datasets of large texts. There are various methods followed for topic modelling.These methods also have variance in accuracy depending on the datasets and its type. Latent Dirichlet Allocation (LDA) has been used for a long time for topic modeling. However, other topic modeling techniques like Dynamic Topic Modeling (DTM), Structural Topic Modeling (STM), and Embedding Topic Modeling (ETM) are also being used which give insightful results. This paper presents a comparative analysis of how these topic modeling methods perform and what factors are responsible for variance in results.**

***Index Terms*—**

I. Introduction

Analyzing and understanding the temporal evolution of topics within this vast collection of articles is a daunting task as the sheer volume of news articles being published every day is staggering. For better understanding of news coverage evaluation, impacts of news articles changes, also for better insights of important news trends we need to temporal analyze the news articles and Dynamic Topic Modeling (DTM), Latent Dirichlet Allocation (LDA), Structural Topic Modeling (STM) and Embedding Topic Modeling (ETM) are some effective approach to do this. Each of the model techniques perform well for different kinds of datasets. Different authors have suggested different models in their paper that worked best in their research and dataset . One of the models, DTM, can be used to analyze changes in news articles over time in a large corpus of documents. Many successful applications of DTM are seen on social media, public speech, literature etc. To use DTM for temporal analysis proper understanding of NLP, ML algorithms and programming knowledge is needed. Again, the mapping between the bag-of-words representation and the embedding space is learned by the ETM model using a neural network. A clustering method is then used to organize the word embeddings into topics. In this paper, we want to make a comparative analysis of these different topic modeling methods. Specifically, we aim to make a comparison of these modeling techniques and identify which model technique performs better than the others. We believe that our study will shed light on the dynamic nature of news coverage and provide valuable insights for journalists, policymakers, and researchers.

II. Literature Review

In this paper, Dieng et al. [1] proposes a new method for topic modeling that incorporates the use of embedding spaces. They discussed that traditional topic modeling techniques such as Latent Dirichlet Allocation (LDA), which are based on bag-of-words representation of text documents has some limitations such as the inability to capture the semantic meaning of words and the difficulty in handling out-of-vocabulary words. The proposed method by the authors, called Embedding Topic Model (ETM), addresses these limitations by incorporating the use of word embeddings, which are distributed representations of words in a continuous vector space. The ETM model uses a neural network to learn the mapping between the bag-of-words representation and the embedding space. It then applies a clustering algorithm to group the word embeddings into topics. The paper evaluates the ETM model on several datasets and compares its performance to traditional topic modeling techniques such as LDA and Non-negative Matrix Factorization (NMF). The results show that the ETM model outperforms these traditional techniques in terms of topic coherence and the ability to handle out-of-vocabulary words.

In this paper Qiang et al. [2] presents a comprehensive survey of short text topic modeling techniques, their applications, and performance. The challenges of topic modeling for short texts such as tweets, product reviews, and search queries was discussed in the paper. Short texts have a limited number of words, and the context is often ambiguous, making it difficult to identify the underlying topics. The paper provides an overview of traditional topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). The authors discussed the recent advancements in short text topic modeling techniques that have been proposed to address the challenges of modeling topics from short texts. These techniques include hierarchical topic modeling, latent topic modeling, and deep learning-based topic modeling. The paper also discusses the applications of short text topic modeling in various domains such as social media analysis, recommendation systems, and opinion mining.The authors also evaluate the performance of the different short text topic modeling techniques using various metrics such as topic coherence and topic diversity. The results show that the recent advancements in short text topic modeling techniques have improved the performance of topic modeling for short texts.

In this paper, Jelodar et al. [3] presents a comprehensive survey of Latent Dirichlet Allocation (LDA) and topic modeling techniques, their applications, and variations. The authors described the key components of LDA, such as the generative model and the inference algorithm, and highlighted the advantages and limitations of LDA. They also present a survey of various variations of LDA and topic modeling techniques, such as dynamic topic modeling, non-negative matrix factorization, and Bayesian non-parametric models. The authors evaluate the performance of LDA and topic modeling techniques using various metrics such as perplexity, coherence, and accuracy. The results show that LDA and topic modeling techniques have good performance in many applications, especially in unsupervised learning scenarios.

In this paper, Hong et al. [4] presented an empirical study of topic modeling in Twitter, including an evaluation of Latent Dirichlet Allocation (LDA) on Twitter data. The paper discussed the unique characteristics of Twitter data, such as the short and informal nature of tweets and the use of hashtags and mentions. The authors evaluate the performance of LDA on Twitter data using various metrics such as coherence and topic uniqueness. They compare the performance of LDA with and without preprocessing steps such as removing stop words, stemming, and removing infrequent terms. The results show that LDA performs well on Twitter data, especially with appropriate preprocessing steps. The authors also conducted a qualitative analysis of the topics generated by LDA on Twitter data. They identified several recurring themes, such as news and current events, technology, and sports. They also showed how LDA can be used to identify emerging topics in real-time, such as breaking news events.

In this paper, Albalawi et al. [5] presented a comparative analysis of several topic modeling techniques on short-text data. The authors discussed the challenges of topic modeling on short-text data, such as the lack of context and the sparse nature of the data. They described the key topic modeling techniques, including Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA). The authors evaluated the performance of these techniques on three different datasets of short-text data, including news headlines, tweets, and product reviews. They compare the performance of these techniques using various metrics, such as coherence and perplexity. The results of the study showed that LDA and NMF outperform LSA in terms of topic coherence and diversity. However, LSA performs better than LDA and NMF in terms of perplexity. The authors also showed that the performance of these techniques is highly dependent on the specific characteristics of the dataset. The authors also conducted a qualitative analysis of the topics generated by these techniques on the three datasets. They identify several recurring themes, such as politics and sports in the news headlines dataset, and customer service and product quality in the product reviews dataset.

In this paper, Sendhilkumar et al. [6] suggests a technique for producing word clouds from text documents via topic modeling. The authors explain that topic modeling is a statistical approach used to recognize topics or themes in a group of documents, and they illustrate how they utilized the Latent Dirichlet Allocation (LDA) algorithm to recognize topics in a given set of documents. They then describe their technique for producing word clouds from the identified topics using the term frequency-inverse document frequency (TF-IDF) weighting method to highlight the most crucial words related to each topic. The authors also compared their method to other word cloud generation approaches and found that it was more effective in generating representative word clouds. The paper provides a comprehensive and precise description of the proposed method for creating word clouds using topic modeling and a thorough evaluation, making it valuable to professionals and researchers in the natural language processing and information visualization fields.

In this paper, Halima Banu et al. [7] the authors argue that traditional topic modeling approaches may not be sufficient for capturing the nuances and complexities of trending topics, and thus propose a sub-topic detection model to improve the accuracy and granularity of topic analysis.The paper first provides an overview of traditional topic modeling approaches, highlighting their limitations in identifying sub-topics and capturing temporal changes in topics. The authors then introduce their sub-topic detection model, which incorporates a novel algorithm that considers the frequency, sentiment, and coherence of words in a given text corpus. The proposed approach is evaluated on a dataset of tweets related to the COVID-19 pandemic, and the results demonstrate that the sub-topic detection model outperforms traditional topic modeling approaches in identifying more granular and relevant sub-topics within trending topics. The paper also provides a visual analysis of the identified sub-topics over time, highlighting the evolution and interrelation of different sub-topics within the COVID-19 topic.

In the paper Miao et al. [8] ”Neural Variational Inference for Text Processing,” Miao, Yu, and Blunsom (2016) suggest a new approach to text processing that uses neural variational inference to address the drawbacks of traditional inference algorithms. By training a neural network to learn a probabilistic model of the text corpus, the authors aim to improve efficiency and overcome overfitting. The proposed method involves predicting the parameters of a variational distribution that approximates the true posterior distribution, using a reparameterization trick to allow for backpropagation through stochastic sampling. The resulting neural variational inference algorithm is demonstrated to produce higher quality topic models with better efficiency compared to traditional methods. The paper's analysis of the approach on various text datasets highlights its ability to identify meaningful topics in the corpus, providing a significant contribution to the field of text processing and showcasing the potential of neural network-based probabilistic modeling.

III. Conclusion

Research on News articles analysis is ongoing and constantly evolving. There has been a significant amount of research conducted on this topic in recent years, with many studies employing the different types of model techniques to analyze changes in topics over time. In our research we plan to get a modeling technique as a result which remains stable with the evaluation of news articles over time. This research will help the journalists and policymakers to have a proper understanding and get a good insight of the trending and upcoming news.

References

[1] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, “Topic Modeling in Embedding Spaces,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, Jul. 2020, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00325. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00325/1923074/tacl_a_00325.pdf. [Online]. Available: https://doi.org/10.1162/tacl%5C_a%5C_00325.

[2] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, “Short text topic modeling techniques, applications, and performance: A survey,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1427–1445, 2022. DOI: 10.1109/TKDE.2020.2992485.

[3] A. Zbiciak and T. Markiewicz, “A new extraordinary means of appeal in the polish criminal procedure: The basic principles of a fair trial and a complaint against a cassatory judgment,” en, *Access to Justice in Eastern Europe*, vol. 6, no. 2, pp. 1–18, Mar. 2023.

[4] L. Hong and B. D. Davison, “Empirical study of topic modeling in twitter,” in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA ’10, Washington D.C., District of Columbia: Association for Computing Machinery, 2010, pp. 80–88, ISBN: 9781450302173. DOI: 10.1145/1964858.1964870. [Online]. Available: https://doi.org/10.1145/1964858.1964870.

[5] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using topic modeling methods for short-text data: A comparative analysis,” *Frontiers in Artificial Intelligence*, vol. 3, 2020, ISSN: 2624-8212. DOI: 10.3389/frai.2020.00042. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2020.00042.

[6] S. Sendhilkumar, M. Srivani, and G. S. Mahalakshmi, “Generation of word clouds using document topic models,” in *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, 2017, pp. 306–308. DOI: 10.1109/ICRTCCM.2017.60.

[7] S. Halima Banu and S. Chitrakala, “Trending topic analysis using novel sub topic detection model,” in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2016, pp. 157–161. DOI: 10.1109/AEEICB.2016.7538263.

[8] Y. Miao, L. Yu, and P. Blunsom, *Neural variational inference for text processing*, 2016. arXiv: 1511.06038 [cs.CL].