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), Correlated Topic Modeling (CTM) 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*—Topic Modeleing, Latent Dirichlet Allocation, Embedded Topic Model, Correlated Topic Modeling**

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), Correlated Topic Modeling (CTM), Non-negative Matrix Factorization (NMF), atent Semantic Analysis (LSA) 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 topic modeling method called Embedding Topic Model (ETM). It involves the use of embedded spaces. The authors analyzed that traditional topic modeling techniques such as Latent Dirichlet Allocation (LDA) based on bag-of-word representations of text documents have several limitations. The inability to grasp the meaning of words and the difficulty of removing them from the vocabulary. The proposed method by the authors addresses these limitations by using word embeddings, which are distributed representations of words in a continuous vector space. ETM models use neural networks to learn the mapping between the bag-of-words representation and the embedding space. A clustering algorithm is then applied to group the word embeddings into topics. The authors evaluated the ETM model on multiple datasets and compared its performance with traditional topic modeling techniques such as LDA and Non-negative Matrix Factorization (NMF). According to the results, in terms of topic coherence and ability to handle out-of-vocabulary words., the ETM model outperforms these conventional techniques.

In this paper Qiang et al. [2] presented a comprehensive survey of short text topic modeling techniques, their applications, and performance. Here, the authors have discussed different challenges of topic modeling for short texts such as tweets, product reviews, and search queries. There is limited word count for short texts and it is seen that these texts are often vague in context for which it is difficult to identify the underlying topics. The authors in the paper provided 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. Moreover, it also discusses the applications of short text topic modeling in different domains such as social media analysis, recommendation systems, and opinion mining. Different metrics like topic coherence and topic diversity were used by the authors to evaluate the performance of the different short text topic modeling techniques. 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] presented 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 provide an overview of different variations of LDA and topic modeling techniques, such as dynamic topic modeling, non-negative matrix factorization, and Bayesian non-parametric models. Moreover, dIfferent metrics like perplexity, coherence, and accuracy were used by the authors to evaluate the performance of LDA and topic modeling techniques. According to the results, the authors have concluded 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 empirical studies of topic modeling in Twitter, including evaluating the Latent Dirichlet Allocation (LDA) of Twitter data. Authors have discussed in the paper about the unique characteristics of Twitter data, such as the short and informal nature of tweets and the use of hashtags and mentions. Moreover, different metrics such as coherence and topic uniqueness were used by the authors to evaluate the performance of LDA on Twitter data. 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.

In the paper Bindu et al. [9] the authors have evaluated two models on a question answer dataset based on autism. The two models they have used are Dynamic Topic Modeling (DTM) and Correlated Topic Modeling (CTM). For measuring, Log Likelihood and Perplexity were used. The authors stated that one of the limitations of LDA that is Dirichlet dispersion utilized in it for free presumptions with respect to the topic extents isn't of genuine significance. Authors considered these limitations and two real time dataset algorithms like DTM and CTM are used for the dataset. Author's proposed algorithm evaluates DTM and CTM on the dataset. From analyzing the perplexity values the two model, it is found that CTM performs better than the DTM.

III. Results and Analysis

In the paper [1], the authors have studied the model Embedded Topic Model (ETM) and compared it to other unsupervised document models. Authors have evaluated performance in terms of both prediction accuracy and topic interpretability since a good document model should offer both coherent linguistic patterns and an accurate word distribution. The accuracy was evaluated with log-likelihood and the topic interpretability using a combination of topic coherence and variety. The authors discover that the ETM offers the best forecasts and subjects among the interpretable models. Moreover, they studied the robustness of the different models with stop words. Stop words frequently appear in documents, every learnt subject contains some stop words, which results in poor topic interpretability. The authors have used the 20Newsgroups corpus and the New York Times corpus which has more than 1.8 million articles. After preprocessing, different models were used on this dataset. Besides ETM, the authors have used Latent Dirichlet allocation (LDA) which is a popular used topic model, Neural Variational Document Model (NVDM) which is a multinomial factor model of documents. The authors evaluated two variants of ETM, one is ETM-PWE where the word embeddings are pre-fitted and one in which they are simultaneously learned with the other parameters. Interpretable topics are only provided by ETM and LDA because other matrices are unconstrained and cannot be interpreted. The authors furthermore conducted a quantitative analysis of the models. They evaluate the subjects' quality as well as the model's capacity for prediction. Topic coherence and topic variety are two variables that were combined to determine topic quality. Topic models for documents are easier to interpret when they have a better topic coherence. The authors have found that in almost all scenarios, LDA prediction is the worst. The NVDM provides the better prediction on the New York Times and on the 20NewsGroups, it performs generally better than LDA but worse than for the other approaches. But both variants of the ETM show effectiveness and provide the best predictive performance while keeping interpretable topics. The authors also evaluated a version of the New York Times corpus that contains all stop words and in this evaluation, ETM-PWE performs best in terms of topic quality.

In the paper [2], the authors are trying to use different topic modeling techniques for short texts. They have chosen 9 topic models LDA, GSDMM, LF-DMM, GPU-DMM, GPU-PDMM, BTM, WNTM, SATM and PTM. Now, the authors selected six datasets - SearchSnippets, StackOverflow, Biomedicine, Tweet, GoogleNews, PascalFlickr to verify the model. Each dataset has 12295, 16407, 19448, 2,472, 11109, 4834 numbers of documents in each dataset. All the models were evaluated using different metrics like classification accuracy, clustering, topic coherence, efficiency. For different datasets and different metrics, different models showed best performance like for 'Biomedicine' dataset LDA and DMM showed better performance than other models. Again, while using topic coherence, DMM showed better performance for all datasets than others. The authors later concluded from the results that word embedding-based techniques like LF-DMM, GPU-DMM, and GPU PDMM are the most effective for short text topic modeling.

In the paper [3], the authors have made a survey on Latent Dirichlet allocation (LDA) and different topic modeling techniques and its applications. In order to understand the evolution of research, current trends, and intellectual structure of topic modeling, the authors looked at highly academic works from the year 2003 to 2016 that were based on LDA and discussed topic modeling. The research

papers that these authors have evaluated contains topics like scientific topic discovery, image classification, opinion mining, source code analysis and many more. In most of these papers, LDA modeling techniques were used. After observing different research papers throughout the years, detailed analysis of different models and its output, the authors have suggested that LDA for text mining with topic modeling.

In the paper [4], the authors wanted to effectively train a standard topic model in short text environments. As a dataset, they have used Twitter data from streaming and regular API. The API is a push-style API with varying levels of access that continuously sends a tiny portion of Twitter messages across an ongoing TCP connection. The authors have worked on ”Garden-hose” level, which the firm defines as offering a ”statistically significant sample” of the messages that are sent through its system. From this they got data of 1,992,758 messages and 514,130 users. Moreover, from twitter suggestions,they crawled users and messages. Thus another dataset contains 52,606 distinct terms and 50,447 messages. Authors have used two models which were Latent Dirichlet allocation (LDA) and Author -Topic (AT) model. They used TF-IDF weighting scores and USER scheme as features and trained a logistic regression classifier. Furthermore, the authors demonstrated through the tests that training a conventional LDA model on user aggregated profiles is superior to the simple addition to the AT model which is an extension of LDA in terms of modeling for messages and users.

In the paper [6], the authors used Latent Dirichlet Allocation (LDA) algorithm, which is a popular probabilistic topic modeling technique, in order to pinpoint topic in the dataset. The proposed method then used frequency-inverse document frequency (TF-IDF) weighting method to create a sample word cloud for each topic and highlight the key terms connected to each topic. 1500 news stories from the BBC website were used to generate the dataset which was used in this research. The authors ultimately compared their approach against other approaches and discovered that LDA outperformed them in terms of visual similarity and semantic coherence.

In the paper [7], the author suggested a sub-topic detection model for trending topic analysis. The authors claim that typical topic modeling techniques may not be enough for capturing the nuances and complexities of trending topics, and hence they presented a sub-topic detection model to improve topic analysis accuracy and granularity. The suggested method includes a novel algorithm that takes into account the frequency, sentiment, and coherence of terms in a text corpus. The dataset that was used consists of tweets related COVID-19 pandemic. The authors implemented the COVID-19 pandemic dataset to demonstrate the usefulness of their proposed sub-topic detection model in order to detect more granular and important sub-topics from trending topics. The research results indicate that in detecting more glandular and relevant sub-topics inside trending subjects, the sub-topic detection model surpasses conventional topic modeling approaches. The study includes a temporal visual analysis of the subtopics that have been found, emphasizing how they have changed and interacted with one another within the COVID-19 topic. Additionally, the authors demonstrate how their method is a useful tool for comprehending complicated and quickly changing topics by demonstrating how well it can capture sentiment and coherence in subtopics.

In the paper [8], the authors in order to improve the efficiency and overcome overfitting in text processing, that are common issues with conventional inference algorithms, used neural variational inference approach. They used a neural network to learn a probabilistic model of the text corpus, that allowed more efficient and accurate identification of meaningful topics in the corpus. The authors evaluated their approach to the 20 Newsgroups dataset as well as Reuters, and NIPS conference papers datasets. These datasets were used by the authors to show how well their method identified important topics in various kinds of text data. The 20 Newsgroups dataset includes messages from 20 distinct newsgroups, offering a wide range of important topics. The Reuters dataset incorporates news stories from many topics, offering a broader and more complicated dataset. The NIPS conference papers dataset contains papers from the Neural Information Processing Systems conference, providing a dataset of technical papers. All in all, the results using the neural variational inference algorithm showed that the suggested method is better than the traditional methods in terms of identifying the meaningful and important topics in the text corpus. Moreover, it has shown the potential of neural network-based probabilistic modeling and provided a notable contribution to the field of text processing.

In the paper [9], the authors used model Dynamic Topic Modeling (DTM) and Correlated Topic Modeling (CTM) on a Question Answer dataset based on autism. The dataset contains files estimating around 4000. Their algorithm had evaluated both CTM and DTM on the dataset based on the measures Log Likelihood and Perplexity. After taking different ratios of dataset for testing and training, the better model was evaluated based on the measures. Based on the results by log likelihood and perplexity values of the model, authors found CTM performing better than the DTM.

After detailed analysis, we have concluded that for different kinds of datasets, different kinds of text topic models perform better. We have found that datasets which have long texts like articles, Embedded Topic Model (ETM) performs best. Again for long texts like Question Answer datasets, Correlated Topic Modeling (CTM) shows better performance. Moreover, The Latent Dirichlet Allocation (LDA) algorithm, despite being a popular text modeling technique for long text dataset, performs better for short text dataset also like tweets, search snippets,

google news etc. CTM, LF-DMM, GPU-DMM, and GPU PDMM are also effective for different datasets of short text topic modeling.

IV. 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].

[9] B. K. R., G. H. G. Krishna, and L. Parameswaran, "A performance evaluation of correlated and dynamic topic modeling on a qa dataset," in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2019, pp. 1–7. DOI: 10.1109/ICECA.2019.8822023.