Enhancing Topic Modeling Through Embedding Learning Strategies

Pallabi Biswas[1], Dipankar Bala[2], Lubna Yasmin Pinky[3], Mohammad Ashraful Islam[1]
Email: pallabiju27@gmail.com, dipubala466@gmail.com, lubnacse@mbstu.ac.bd, ashraful.islam@juniv.edu
[1]Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh.
[2]Department of Software Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh
[3]Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University
Tangail, Bangladesh.

Abstract*—In the field of Natural Language Processing (NLP), topic modeling is crucial for uncovering patterns in textual data. Recent advances have combined traditional topic modeling with word embeddings, introducing the Embedded Topic Model (ETM). This paper explores embedding learning strategies within topic modeling to improve the ETM and related models. It delves into more efficient variational inference, advanced word embedding techniques, and strategies for better topic interpretability. Practical implications in document classification, content recommendation, and summarization are evaluated. Scalability challenges for handling large textual corpora are also addressed. The integration of textual data with other modalities is pioneered. This work aims to enhance topic modeling using embedding learning strategies, bridging the gap between theory and practice in NLP.

Index Terms*—Topic Modeling, Embedded Topic Model (ETM), Word Embeddings, Variational Inference, Topic Coherence, Semantic Similarity.

I. INTRODUCTION

In the dynamic landscape of Natural Language Processing (NLP), topic modeling stands as a critical pillar for uncovering insights within vast text datasets. Conventional techniques such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) have served well, the exponential growth of textual data presents new challenges. Recent advances have introduced Word embeddings, which encapsulate the meaning and contextual relationships of words in continuous vector spaces. The Embedded Topic Model (ETM) excels in extracting interpretable topics even from extensive vocabularies, including rare and stop words [6]. This paper outlines a mission to "Enhance Topic Modeling Through Embedding Learning Strategies." The research focuses on refining embedding-based topic modeling and exploring strategies to maximize its potential. The development of efficient variational inference algorithms, employment of advanced word embedding techniques, and pioneering innovative methods for enhancing topic interpretability are key goals. The performance of the ETM in real-world scenarios, including document classification, content recommendation, and topic-based summarization, is evaluated. Scalability challenges in handling large datasets are also addressed. The integration of multimodal data, where text is seamlessly combined with other data types, is investigated, expanding the possibilities of embedding-based topic modeling.

II. LITERATURE REVIEW

The literature on advanced machine learning models for topic modeling highlights a range of innovative approaches but also presents clear limitations in scalability and adaptability to different contexts. One of the main areas of focus has been entity relationship extraction, where models such as the Bidirectional LSTM have demonstrated effectiveness in capturing semantic relationships within text [12]. However, these models tend to require significant computational resources, which can hinder their applicability in resource-constrained environments. Similarly, knowledge graph construction from crime reports shows promise in organizing and analyzing criminal data, but it often overlooks the differences in resource availability across various law enforcement agencies [9].

Research on sarcasm detection in online social media platforms has employed models like Techniques like Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) networks [1]. While these models offer robust solutions, they are limited by the challenges of detecting sarcasm across different cultural contexts and social media platforms. For instance, sarcasm can vary significantly depending on language and the platform, making it difficult for the models to generalize effectively.

In addition to these, topic modeling has been explored for more specific applications such as project idea similarity and document similarity. Keyword extraction and general models like Latent Dirichlet Allocation (LDA) [2] and [4] and Gibbs-EM are often used, but the depth of topic interpretability remains a concern. The complexity of topics in real-world applications demands models that can go beyond surface-level keyword matching to more nuanced topic representations. The detection of suicidal ideation through Twitter data is another example where machine learning techniques are applied to the crucial direction. Methods such as clustering, natural language processing (NLP), and association rules have been used to classify and analyze content related to mental health [8]. However, the ambiguity in labeling suicidal ideation poses challenges, and the models may need further refinement to improve detection accuracy.

In recent studies, an evolving emphasis on improving the efficiency and interpretability of topic models. For example, the Homologous Automated Document Exploration and Summarization (HADES) system, which integrates LDA and

NMF, is effective in many cases but struggles with handling unstructured and multi-topic documents [14]. One popular method is to transform the discrete text into continuous embedding observations, and then modify LDA to produce real-valued data [5]. Similarly, parsimonious topic models that focus on salient word discovery face issues of computational scalability and sparse representation [13].

III. Background

LDA and word embeddings are the foundational concepts that the ETM works upon. Take a look at a corpus of D papers with V unique terms in the lexicon. Let $w_{dn} \in \{1, \ldots, V\}$ represents the n-th word in the d-th document.

A. Latent Dirichlet allocation

Documents are mixes of topics, and topics are distributions over words, according to the fundamental topic modeling technique known as LDA. It is a probabilistic generative model for documents [2]. The generative process includes the following steps:

Draw topic proportions θ_d for each document d from a Dirichlet prior Dir(α).

For each word w_n in document d:

Select a topic z_n based on θ_d.

Draw w_n from the topic's word distribution ϕ_{zn}.

The combined likelihood of latent variables and words is

$$P(w, z, \theta, \phi \mid \alpha, \beta) = \prod_{d=1}^{D} P(\theta_d \mid \alpha) \prod_{n=1}^{N_d} P(z_{dn} \mid \theta_d) P(w_{dn} \mid \phi_{z_{dn}}) P(\phi \mid \beta) \quad (1)$$

Equation (1) represents the joint probability distribution of words w, topic assignments z, document-topic distributions θ, and topic-word distributions ϕ, given the hyperparameters α and β.

B. Word2Vec

Word2Vec comprises a set of models designed to generate word embeddings. These models utilize shallow, two-layer neural networks trained to predict the surrounding context of words. The primary models include:

(a) Continuous Bag-of-Words (CBOW): Predicts the target word from the context words.
(b) Skip-Gram: Predicts the context words from the target word.

These embeddings capture semantic relationships between words, allowing similar words to be placed close to each other in the vector space.

C. Embedded Topic Model (ETM)

ETM combines LDA with word embeddings. It uses continuous representations of words to model the topic distributions, enhancing the model's ability to capture semantic relationships and improving topic coherence.

IV. Proposed System

The proposed system seeks to address several critical limitations in existing Embedded Topic Models (ETM), particularly focusing on enhancing accessibility and usability for a broader audience. While current ETM implementations often require significant expertise to operate effectively, there is a noticeable gap in user-friendly tools and interfaces that could simplify their implementation and customization.

Hyperparameter	Value
Optimizer	Adam
batch_size	512
Epochs	1000
theta_act	leakyrelu
eval_batch_size	1000
num_topics	50
t_hidden_size	1048
lr	0.002
enc_drop	0.4
clip	0.5
wdecay	$1e-5$
additional_hidden_size	600
additional_activation	selu

TABLE I: Hyperparameter Used In The Model

This complexity acts as a barrier, preventing non-experts and beginners from leveraging the power of ETM in their respective domains. Moreover, existing models struggle with challenges such as suboptimal initialization strategies, ineffective management of KL divergence during training, and insufficient regularization in encoder networks. These issues contribute to less accurate topic representations and a lack of distinct, interpretable topic embeddings, ultimately reducing the coherence and utility of the model's outputs. The lack of robust dropout methods and inadequate optimization strategies further exacerbate these problems, often leading to overfitting and suboptimal performance.

As shown in TABLE I some important hyperparameters used in the model. By developing more accessible tools alongside improving the underlying model architecture and training processes, this work aims to facilitate the wider adoption of ETM across various fields, enabling both experts and novices to effectively apply topic modeling to their data.

A. Improved Algorithm

Here is the improved algorithm. The core strategies are unchanged, which is taken from "Topic Modeling in Embedding Spaces [6]

Initialize model parameters $\alpha_{1:K}$ using Xavier initialization.

Initialize variational parameters v_μ and v_Σ.

for iteration $i = 1, 2, \ldots,$ **max_iterations do**

Compute $\beta_k = \text{softmax}(\rho^\top \alpha_k)$ for each topic k

Choose a minibatch B of documents
For each document d in B do
- Get normalized BoW representation x_d
- Compute $\mu_d = NeuralNetwork\ (x_d, v_\mu)$
- Compute $\sum_d = NeuralNetwork\ (x_d, v_\Sigma\ )$ with constraints for positive semi-definiteness
- Sample $\theta_d \sim logNormal(\mu_d, \sum_d\ )$using reparameterization trick
- for each word w_{dn} in document d do
  Compute $P(w_{dn}|\theta_d) = \theta_d^T \beta\ .w_{dn}$
- end for

end for
- Estimate the ELBO using Monte Carlo approximation
- Compute gradient of ELBO w.r.t. model parameters α1: K and variational parameters (νμ, νΣ) using backpropagation
- Update model parameters α1:K using Adam optimizer with decaying learning rate and weight decay regularization
- Update variational parameters (νμ, νΣ) with gradient ascent and gradient clipping

end for

B. Algorithm Improvements

We enhanced the Embedded Topic Model (ETM) incorporate several key improvements aimed at increasing model performance, stability. Firstly, we utilize Xavier initialization for model parameters $\boldsymbol{\alpha_{1:K}}$ to ensure better convergence. We also implement KL annealing, which gradually increases the weight of the KL divergence term during the initial epochs, striking a balance between learning and regularization. The encoder network ($q\theta$) has been augmented with additional layers and batch normalization, resulting in faster convergence and reduced overfitting. Additionally, we apply variational dropout across the entire network, rather than limiting it to the encoder output, for more comprehensive regularization. To mitigate overfitting, early stopping is applied based on validation loss. To ensure the integrity of the covariance matrices, we enforce constraints for positive semi-definiteness when computing Σ_d. For updating model parameters $\alpha_{1:K}$, we use the Adam optimizer with a decaying learning rate and weight decay regularization, which contributes to stable learning. Finally, we incorporate gradient clipping during the update of variational parameters (ν_μ, ν_Σ) to mitigate issues associated with exploding gradients. These modifications collectively enhance the ETM, leading to faster convergence, reduced overfitting, and more reliable topic representations.

C. Methodology

The methodology covers data collection, preprocessing, dataset splitting, training, topic modeling, accuracy measures, testing, and result analysis.

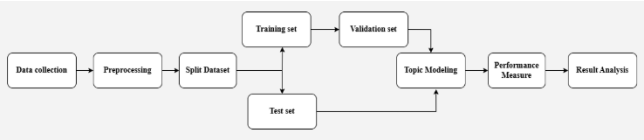

Fig. 1: Topic modeling through parameter initialization (Model-1)

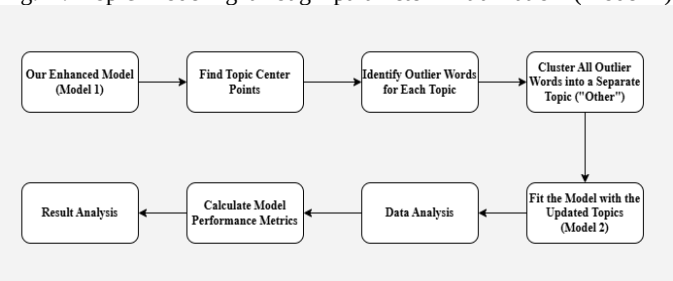

Fig. 2: Improved Model via outlier clustering (Model-2)

D. Model

1) Generative Process:

ETM integrates word embeddings into the topic modeling framework. The generative process for ETM is [6]:

- For each topic *k*, draw a topic embedding β_k from a Gaussian prior $N(0, I)$.
- For each document *d*:
  - Draw a topic assignment z_{dn} from θ_d.
  - Draw wdn from a softmax distribution

parameterized by the topic embedding βzdn and the word embedding matrix ρ.

Formally, the generative process is described as:

$$\beta_k \sim \mathrm{N}(0, I) \quad (2)$$

$$\theta_d \sim \mathrm{Dir}(\alpha) \quad (3)$$

$$z_{dn} \sim \mathrm{Categorical}(\theta_d) \quad (4)$$

$$w_{dn} \sim \mathrm{Softmax}(\rho \cdot \beta_z dn) \quad (5)$$

Equation (2) states that the topic vectors β_k are drawn from a multivariate normal distribution with a mean of zero and an identity covariance matrix *I*.

Equation (3) the topic distribution θ_d for document *d* is sampled from a Dirichlet distribution with parameter α. The parameter α controls the sparsity of the topic distribution, with higher values leading to more uniform distributions over topics.

Equation (4) z_{dn} represents the topic assignment for word *n* in document *d*. This assignment is drawn from a categorical distribution determined by θ_d, the topic distribution for the document. Each word is assigned a topic according to this distribution.

Equation (5) w_{dn} denotes the word corresponding to word *n* in document *d*. The word is generated by sampling from a Softmax distribution, where the input to the Softmax function is the topic-specific vector $\beta_z dn$, scaled by a factor ρ. The SoftMax function transforms the topic vector into a probability distribution over words, which is used to select the word w_{dn}.

2) Inference:

Variational inference is used to approximate the posterior distribution of latent variables. The variational distribution for ETM is:

$$q(\theta_d, z_d) = q(\theta_d \mid \gamma_d) \prod_{\{n=1\}}^{\{N_d\}} q\left(z_{\{dn\}} \mid \phi dn\right) \quad (6)$$

Equation (6) defines the variational approximation $q(\theta_d, z_d)$ for the joint distribution of the topic distribution θ_d and topic assignments z_d for document d. where γ_d and ϕ_{dn} are variational parameters for the topic proportions and topic assignments, respectively. The evidence lower bound (ELBO) is maximized to optimize these parameters:

$$L = E_q[\log P(w, \theta, z \mid \alpha, \beta)] - E_q[\log q(\theta, z)] \quad (7)$$

Equation (7) defines the Evidence Lower Bound (ELBO) L. It represents the difference between the expected log- likelihood of the observed data under the model $P(w, \theta, z \mid \alpha, \beta)$ and the expected log of the variational distribution $q(\theta, z)$. Maximizing L helps approximate the true posterior distribution.

V. Results and Data Analysis

A. Dataset

We study the 20Newsgroups corpus, a benchmark dataset for this work. This dataset consists of a collection of newsgroup documents. The 20 Newsgroups dataset is widely used for experiments in text-based applications. The dataset is available on bitbucket. We preprocess the corpus by tokenizing, removing stop words, and removing terms with high document frequency.

The corpus was divided into three sets: 100 documents for validation, 7,532 documents for testing, and 11,214 documents for training. TABLE II shows number of total tokens and TABLE III depicts statistics of document frequency(df).

Vocabulary size: 3072

Corpus Sparsity: 0.9825 (98.25%)

TABLE II: Total Token Statistics

Dataset Split	Topic
Train	603,211
Validation	5134
Test	397,790

TABLE III: Document frequency statistics

Document Frequency Statistics	Value
Number of terms with non-zero df	3,072
Average document frequency	196.36
Maximum document frequency	5,705
Minimum document frequency	25

B. Metrics

We use the following metrics for evaluation:

Topic Coherence: Measures the semantic similarity among the top words in a topic, indicating the interpretability of the topics

Topic Diversity: Assesses the distinctiveness of topics within a model, ensuring that the topics cover a wide range of themes.

C. Baselines

(a) **Latent Dirichlet Allocation (LDA):** An established topic model with distinct word representations.

(b) **Embedded Topic Model (ETM):** A model that combines word embeddings with topic models

(c) **Our Model**: Our proposed model that improves word embeddings with topic models.

D. Data Analysis

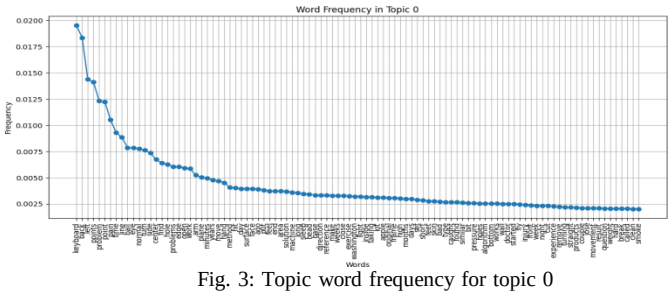

Fig. 3: Topic word frequency for topic 0

E. Result

bar chart, line plot and word cloud. Fig. 3 illustrates topic word frequency of first 100 words for topic 0 is represented as a line plot. It contains a large number of words, which overlap with each other, making the figure dense.

Fig. 4 depicts topic word frequency for topic 2 is represented by bar chart. According to their relevance score top 10 words are shown.

As shown in fig. 5 top words of topic 1 to topic 10 is represented by line plot. According to their relevance score words are presented. These statistics are shown in tabular format in TABLE IV. We measure the center of each topic, then

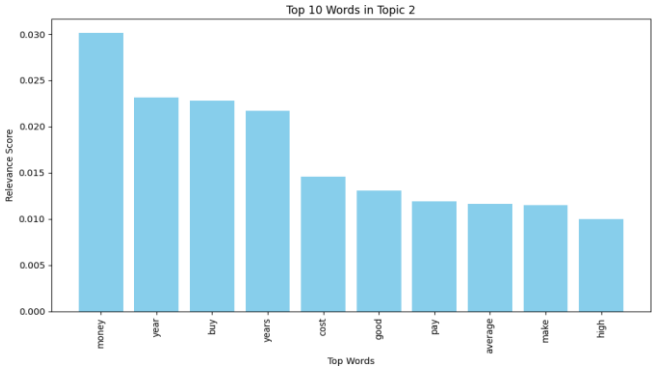

Fig. 4: Top 10 words in topic 2

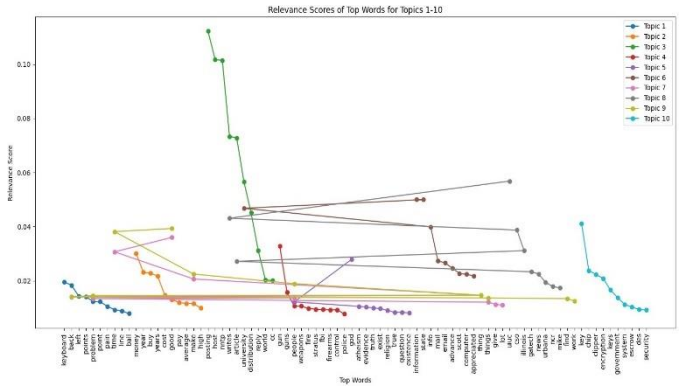

Fig. 5: Top words of topic 1 to 10 (Model-1)

detect the outlier word per topic using distance-based metric (cosine similarity), collect all outlier words from different topics and group them into a single additional topic named "Other" using k-means clustering. After that retrain the model by including the new "Other" topic alongside the original topics. As shown in fig. 6 top words of topic 1 to topic 10 is represented by line plot after retraining model. According to their relevance score words are presented. These statistics are shown in tabular format in TABLE V.

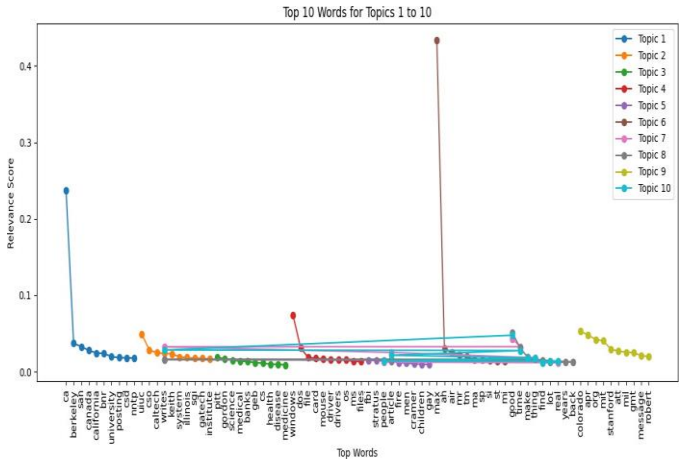

Fig. 6: Top words of topic 1 to 10 after outlier clustering using Model-2

TABLE IV: Top words relevance score

Index	Topic	Word	Relevance Score
0	1	keyboard	0.019526
1	1	back	0.018363
2	1	left	0.014378
3	1	points	0.014127
4	1	problem	0.012327
...	...	...	...
95	10	government	0.013786
96	10	system	0.011216
97	10	escrow	0.010298
98	10	des	0.009332
99	10	security	0.009286

TABLE V: Top words relevance score after re-training (Model-2)

Index	Topic	Word	Relevance Score
0	1	ca	0.236876
1	1	berkeley	0.037235
2	1	san	0.032680
3	1	canada	0.028077
4	1	california	0.024358
...	...	...	...
95	10	make	0.017774
96	10	people	0.013925
97	10	lot	0.013792
98	10	real	0.013676
99	10	find	0.011919

In topic modeling, the topic-word distribution matrix is an essential output that captures the relationship between topics and words in a corpus. Each row in this matrix denotes a topic, and each column denotes a vocabulary word. Higher values indicate stronger correlations between the word and the topic. The values in this matrix represent the probability or relevance of a term given a certain topic. Our model has 50 topics. Topic-word distribution matrix for our model is obtained properly. Fig. 7 illustrates word cloud of topic word distribution matrix of topic 1.

In a similar vein, another essential result of topic modeling is the document-topic distribution matrix which describes the connection between subjects and documents as shown in Fig.8. Each row in this matrix represents a document found in the corpus, and each column denotes a topic. Higher values in the matrix indicate that a topic is more prominent in the corresponding document. The values in the matrix represent the likelihood or relevance of a topic within a document.

As shown in fig.9 Topic 1 and its most relevant words in embedding spaces. The topic is start in the word embedding spaces and about file found by model.

Fig. 7: Word cloud for topic 1 word distribution matrix

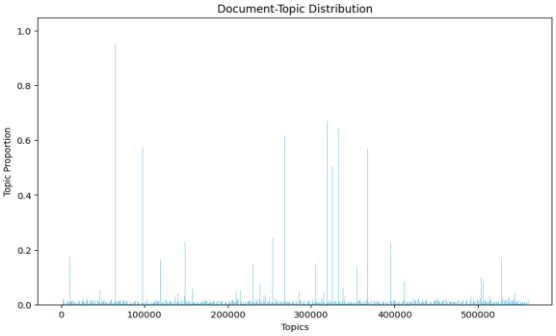

Fig. 8: Document topic distribution line plot

E. Result

The concept of topic coherence states that terms that frequently appear in the same publications will be displayed in a coherent subject. Stated differently, a coherent topic's most likely terms should have a high degree of mutual information. subject models that exhibit greater subject coherence in their documents are easier to understand. Diversity around 0 denotes themes that are redundant, whereas diversity near 1 denotes topics that are more diversified. We define the product of a model's topic coherence and variety as the overall metric for the quality of its topics.

In our model, we have good coherence score and diversity value. They are 0.19830 and 0.658. After re-training the model by identifying and clustering outlier words, we achieved improved results with a coherence score of 0.211 and a diversity value of 0.66.

We have also obtained the most similar word of word dictionary. TABLE VI presents some of them as tabular format. Here, there are ten top words. For every top word in the list, the algorithm yields the ten most related terms. Usually, the result is a dictionary with lists of the most comparable terms as values and the input words as keys. The value 0.001 is the learning rate at which our model's parameters are being updated. A learning rate of 0.001 is quite standard. KL theta represents the Kullback-Leibler divergence for the theta distribution. Generally speaking, lower numbers suggest that the approximation posterior is nearer the genuine posterior. The KL theta of our model provides a suitable value.

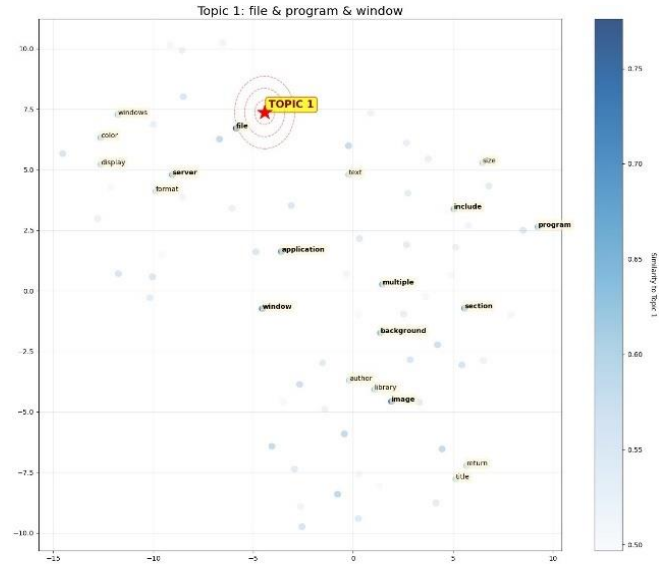

Fig. 9: Topic 1 and its most relevant words in embedding spaces

TABLE VI: Similar Words Table

Data	Analysis	Country	Network	Computer	Student	Food	Exercise	Money	Media
data	analysis	country	network	computer	student	food	exercise	money	media
access	theory	responsibility	engineering	engineering	robert	doctors	injury	pay	report
digital	united	actions	computer	network	familiar	hospital	attacks	years	university
mechanism	basis	education	computers	systems	daily	treatment	factors	extra	statistics
code	educational	military	send	tech	college	doctor	skin	benefit	tim
software	activity	lives	tech	computing	nntp	drugs	paragraph	hear	national
Data	Analysis	Country	Network	Computer	Student	Food	Exercise	Money	Media
technical	primarily	dealing	systems	university	school	health	technique	market	richard
additional	defined	press	university	internet	posting	disease	severe	big	news
included	organizations	crime	local	science	friend	visit	crimes	made	member
limit	foundation	kill	ch	martin	ignore	anti	sexual	lot	Plain

Reconstruction loss is a metric that assesses how successfully the model uses the latent representation to recover the input data. The Reconstruction Loss of our model provides a suitable value. The Negative Evidence Lower Bound (NELBO) is the objective function we are optimizing. Lower values are better, indicating a better fit of the model to the data.

We calculate topic quality by determining two metrics: topic coherence and topic diversity. Topic coherence is a quantitative measure of the interpretability of a topic [11] It is the average point wise mutual information of two words drawn randomly from the same document [7]

$$TC = \sum_{k=1}^{K} \frac{1}{50} \sum_{i=1}^{10} \sum_{j=i+1}^{10} f(w_i^{(k)}, w_j^{(k)}) \frac{1}{K} \quad (8)$$

Equation (8) defines the Topic Coherence (TC) metric, which evaluates the quality of topics in a topic model. It measures the semantic similarity between the top words in each topic by computing a pairwise function f (w(k), w(k)) over the top 10 words of each topic k. The normalization factors ensure an averaged coherence score across all topics. Higher TC values indicate more coherent and meaningful topics.

$$f(w_i, w_j) = \frac{log \frac{P(w_i, w_j)}{P(w_i)\, P(w_j)}}{-log P(w_i, w_j)} \quad (9)$$

Equation (9) defines the coherence function $f(w_i, w_j)$, which quantifies the semantic similarity between two words w_i and w_j. The quantity $P(w_i, w_j)$ is the probability of words w_i and w_j co-occurring in a document, and P (w_i) is the marginal probability of word w_i. We obtain these probabilities with empirical counts.

The concept of topic coherence states that terms that frequently appear in the same publications will be displayed in a coherent subject. Stated differently, a coherent topic's most likely terms should have a high degree of mutual information. subject models that exhibit greater subject coherence in their documents are easier to understand.

We pair a method diversity, a second metric, with coherence. Diversity around 0 denotes themes that are redundant, whereas diversity near 1 denotes topics that are more diversified. We define the product of a model's topic coherence and variety as the overall metric for the quality of its topics.

The Topic Composite Quality (TCQ) metric is a holistic measure designed to evaluate the overall quality of topics generated by topic models. This holistic approach helps in fine-tuning models for better overall performance in real-world applications. The performance comparison of different models and ours is presented in TABLE VII, showing the coherence, diversity, and quality metrics.

TABLE VII: Comparison Table

Model	Coherence	Diversity	Quality
LDA [3]	0.13	0.14	0.0173
-NVDM [10]	0.17	0.11	0.0187
Labeled ETM [6]	0.18	0.22	0.0405
Model-1 (With Parameter Initialization)	0.198	0.65	0.428
Model-2 (Parameter Initialization with outlier clustering)	0.211	0.66	0.44

As shown in fig. 10 heatmap metrics of Enhanced ETM model before and after retraining. It can be said that after retraining the model performs better and provides better performance metrics.

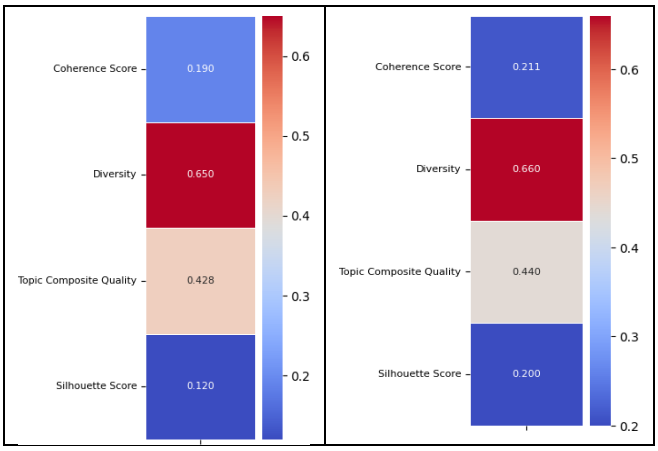

Fig. 10: Comparison of heatmaps Model-1 vs Model-2

VI. Conclusions and Future Work

A. Conclusion

We have obtained an improved topic model that integrates word embeddings into the probabilistic topic modeling framework. Even in datasets with extensive vocabulary sets, the model is able to learn interpretable word embeddings and topics. We looked at how well the model performed in comparison to various document models. The model picks up precise word distribution as well as logical linguistic patterns Based on coherence metrics and diversity this model performs well.

B. Future Work

Bioinformatics: Current topic modeling approaches in bioinformatics lack optimization for specific data types, such as microarrays. Future work could focus on developing tailored models that better capture the unique structure and complexity of biological datasets. With the rapid growth of biological data, enhanced models could provide deeper insights into hidden biological patterns and improve the interpretation of complex biological phenomena.

Summarization: In the realm of opinion and meeting summarization, it can be further improved by integrating more advanced sentence-level opinion detection and event correlation techniques. Additionally, refining the use of topic modeling with speech recognition in automatic meeting summaries could enhance content recall and segmentation accuracy, reducing the need for manual intervention while increasing efficiency.

Real-World Application Evaluation: Conduct studies on the effectiveness of ETM in real-world scenarios, such as social media analysis, customer feedback analysis, or other niche areas. This could involve integrating ETM with other modalities (e.g., images, structured data) to explore its performance in multi-modal contexts.

REFERENCES

[1] Aruna Bhat and Govind Narayan Jha. “Sarcasm detection of textual data on online Social Media: a review”. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE. 2022

[2] David M Blei. “Probabilistic topic models”. In: vol. 55. 4. ACM New York, NY, USA, 2012.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3 Jan (2003).

[4] Stefan Bunk and Ralf Krestel. “Welda: Enhancing topic models by incorporating local word context”. In: *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*. 2018.

[5] Rajarshi Das, Manzil Zaheer, and Chris Dyer. “Gaussian LDA for topic models with word embeddings”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015.

[6] Adji B Dieng, Francisco JR Ruiz, and David M Blei. “Topic modeling in embedding spaces”. In: vol. 8. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . ., 2020, pp. 439–453.

[7] Quoc Le and Tomas Mikolov. “Distributed representations of sentences and documents”. In: *International conference on machine learning*. PMLR. 2014.

[8] Prabhakar Marry et al. “Suicidal Ideation Detection: Application of Machine Learning Techniques on Twitter Data”. In: *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*. IEEE. 2023, pp. 345–350.

[9] Svitlana Mazepa et al. “Relationships Knowledge Graphs Construction Between Evidence Based on Crime Reports”. In: (2022).

[10] Yishu Miao, Lei Yu, and Phil Blunsom. “Neural variational inference for text processing”. In: (2016).

[11] David Mimno et al. “Optimizing semantic coherence in topic models”. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, pp. 262–272.

[12] Minyong Shi, Jingyi Huang, and Chunfang Li. “Entity relationship extraction based on BLSTM model”. In: *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. IEEE. 2019, pp. 266–269.

[13] Hossein Soleimani and David J Miller. “Parsimonious topic models with salient word discovery”. In: vol. 27. 3. IEEE, 2014, pp. 824–837.

[14] Piotr Wilczyn´ski et al. “HADES: Homologous Auto- mated Document Exploration and Summarization”. In: 2023.

Biography

Pallabi Biswas received her Master of Science (Eng.) degree in Computer Science and Engineering from Jahangirnagar University, Bangladesh. She also completed her Bachelor of Science (Eng.) in Computer Science and Engineering from the same institution in June 2023. Her academic journey has provided her with a strong foundation in computational research, and she is particularly passionate about artificial intelligence, Natural language processing, deep learning, and data science. Email: pallabiju27@gmail.com

Dipankar Bala completed his graduation in Software Engineering dept. from Shahjalal University of Science and Technology (SUST), Sylhet, in 2023. He is currently working as a Software Engineer at Brain Station 23. His research interests include Machine Learning, Natural language processing, Artificial Intelligence, and Computer Vision. He is passionate about developing intelligent systems and exploring innovative solutions in these fields. Contact at dipubala466@gmail.com.

Lubna Yasmin Pinky is an Assistant Professor in the Department of Computer Science and Engineering at Mawlana Bhashani Science and Technology University, Bangladesh. Her research interests encompass Natural Language Processing (NLP), Artificial Intelligence (AI), Deep Learning, Social Media Data Analysis, and Recommender Systems. She is passionate about developing intelligent computational models and leveraging data-driven approaches to enhance decision-making and automation in various domains.

Mohammad Ashraful Islam is an Assistant Professor in the Department of Computer Science and Engineering at Jahangirnagar University, Savar, Dhaka, Bangladesh. His research interests include machine learning, artificial intelligence, and data science. With years of academic and research experience, he has contributed to various scholarly publications and has been actively involved in mentoring students. He can be reached at ashraful.islam@juniv.edu.