

Interpretable Topic Analysis

Mincheol Kim

Data Science, Swiss Institute of Artificial Intelligence

September 22, 2023

Abstract

User-generated data, often characterized by its brevity, informality, and noise, poses a significant challenge for conventional natural language processing techniques, including topic modeling. User-generated data encompasses informal chat conversations, Twitter posts laden with abbreviations and hashtags, and an excessive use of profanity and colloquialisms. Moreover, it often contains "noise" in the form of URLs, emojis, and other forms of pseudo-text that hinder traditional natural language processing techniques.

This study sets out to find a principled approach to objectively identifying and presenting improved topics in short, messy texts. Topics, the thematic underpinnings of textual content, are often "hidden" within the vast sea of user-generated data and remain "undiscovered" by statistical methods, such as topic modeling.

We explore innovative methods, building upon existing work, to unveil latent topics in user-generated content. The techniques under examination include Latent Dirichlet Allocation (LDA), Reconstructed LDA (RO-LDA), Gaussian Mixture Models (GMM) for distributed word representations, and Neural Probabilistic Topic Modeling (NPTM).

Our findings suggest that NPTM exhibits a notable capability to extract coherent topics from short and noisy textual data, surpassing the performance of LDA and RO-LDA. Conversely, GMM struggled to yield meaningful results. It is important to note that the results for NPTM are less conclusive due to its extended computational runtime, limiting the sample size for rigorous statistical testing.

This study addresses the task of objectively extracting meaningful topics from such data through a comparative analysis of novel approaches. Also, this research contributes to the ongoing efforts to enhance topic modeling methodologies for challenging user-generated content, shedding light on promising directions for future investigations.

This study presents a comprehensive methodology employing Graphical Neural Topic Models (GNTM) for textual data analysis. "Group information" here refers to topic proportions (θ). We applied a Non-Linear Factor Analysis (FA) approach to extract this intricate structure from text data, similar to traditional FA methods for numerical data. Our research showcases GNTM's effectiveness in uncovering hidden patterns within large text corpora, with attention to noise mitigation and computational efficiency. Optimizing topic numbers via AIC and agglomerative clustering reveals insights within reduced topic sub-networks.

Future research aims to bolster GNTM's noise handling and explore cross-domain applications, advancing textual data analysis.

1 Introduction

Over the past few years, the volume of news information on the Internet has seen exponential growth. With news consumption diversifying across various platforms beyond traditional media, topic modeling has emerged as a vital methodology for analyzing this ever-expanding pool of textual data. This introduction provides an overview of the field and the seminal work of foundations.

1.1 Seminal Work: Topic Modeling Research

One of the pioneering papers in news data analysis using topic modeling is "Latent Dirichlet Allocation", that is, LDA technique, which revolutionized the extraction and analysis of topics from textual

data. The need for effective topic modeling in the context of the rapidly growing user-generated data landscape has been emphasized. The challenges posed by short, informal, and noisy text data, including news articles, are highlighted.

There are numerous advantages of employing topic modeling techniques for news data analysis, including:

- Topic derivation for understanding frequent news coverage.
- Trend analysis for tracking news trends over time.
- Identifying correlations between news topics.
- Automated information extraction and categorization.
- Deriving valuable insights for decision-making.

Recent advancements in the fusion of neural networks with traditional topic modeling techniques have propelled the field forward. Papers such as "Neural Topic Modeling with Continuous Neighbors" have introduced innovative approaches that warrant exploration. By harnessing deep learning and neural networks, these approaches aim to enhance the accuracy and interpretability of topic modeling.

Despite the growing importance of topic modeling, existing topic modeling methods do not sufficiently consider the context between words, which can lead to difficult interpretation or inaccurate results. This limits the usability of topic modeling. The continuous expansion of text documents, especially news data, underscores the urgency of exploring its potential across various fields. Public institutions and enterprises are actively seeking innovative services based on their data.

To address the limitations of traditional topic modeling methods, this paper proposes the Graphical Neural Topic Model (GNTM). GNTM integrates graph-based neural networks to account for word dependencies and context, leading to more interpretable and accurate topics.

1.2 Research Objectives

This study aims to achieve the following objectives:

- Present a novel methodology for topic extraction from textual data using GNTM.
- Explore the potential applications of GNTM in information retrieval, text summarization, and document classification.
- Propose a topic clustering technique based on GNTM for grouping related documents.

In short, the primary objectives are to present GNTM's capabilities, explore its applications in information retrieval, text summarization, document classification, and propose a topic clustering technique.

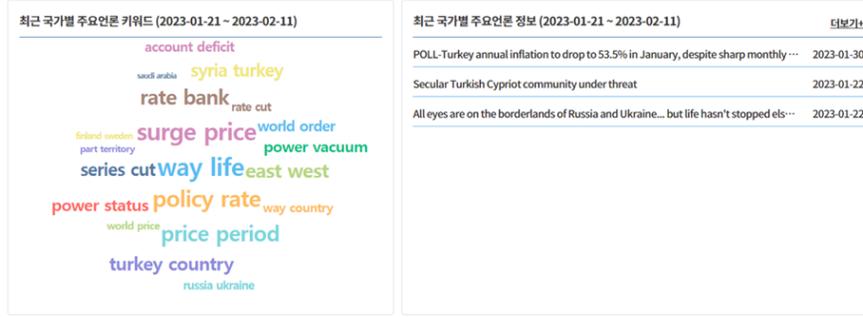
The subsequent sections of this thesis delve deeper into the methodology of GNTM, experimental results, and the potential applications in various domains. By the conclusion of this research, these contributions are expected to provide valuable insights into the efficient management and interpretation of voluminous document data in an ever-evolving information landscape.

2 Problem Definition

2.1 Existing Industry-specific Keywords Analysis

South Korea boasts one of the world's leading economies, yet its reliance on foreign demand surpasses that of domestic demand, rendering it intricately interconnected with global economic conditions[3]. This structural dependency implies that even a minor downturn in foreign economies could trigger a recession within Korea if the demand for imports from developed nations declines. In response, public organizations have been established to facilitate Korean company exports worldwide.

However, the efficacy of these services remains questionable, with South Korea's exports showing a persistent downward trajectory and a trade deficit anticipated for 2022. The central issue lies in the inefficient handling of global textual data, impeding interpretation and practical application.



(a) Country-specific keywords



(b) Industry-specific keywords

Figure 1: Data service provided by public organizations

Han, G.J.(2022) scrutinized the additional features and services available to paid members through the utilization of big data and AI capabilities based on domestic logistics data[5]: Trade and Investment Big Data (KOTRA), Korea Trade Statistics Information Portal (KTSI), GoBiz Korea (SME Venture Corporation), and K-STAT (Korea Trade Association).

Regrettably, these services predominantly offer basic frequency counts, falling short of delivering valuable insights. Furthermore, they are confined to providing internal and external statistics, rendering their output less practical. While BERT and GPT have emerged as potential solutions, these models excel in generating coherent sentences rather than identifying representative topics based on company and market data and quantifying the distribution of these topics.

2.2 Proposed Model for Textual Data Handling

To address the challenge of processing extensive textual data, we introduce a model with distinct characteristics:

1. Extraction of information from data collected within defined timeframes.
2. A model structure producing interpretable outcomes with traceable computational pathways.
3. Recommendations based on the extracted information.

Previous research mainly relied on basic statistics to understand text data. However, these methods have limitations, such as difficulty in determining important topics and handling large text sets, making it hard for businesses to make decisions.

Our research introduces a method for the precise extraction and interpretation of textual data meaning via a natural language processing model. Beyond topic extraction, the model will uncover interrelationships between topics, enhance text data handling efficiency, and furnish detailed topic-related insights. This innovative approach promises to more accurately capture the essence of textual data, empowering companies to formulate superior strategies and make informed decisions.

2.3 Scope and Contribution

This study concentrates on the extraction and clustering of topics from textual data derived from numerous companies' news data sources.

However, its scope is confined to outlining the methodology for collecting news data from individual firms, extracting topic proportions, and clustering based on these proportions. We explicitly state the study's limitations concerning the specific topics under investigation to bolster the research's credibility. For instance, we may refrain from delving deeply into a particular topic and clarify the constraints on the generalizability of our findings.

The proposed methodology in this study holds the potential to facilitate the effective handling and utilization of this vast text data reservoir. Furthermore, if this methodology is applied to Korean exporters, it could play a pivotal role in transforming existing export support services and mitigating the recent trade deficit.

3 Literature Review

3.1 Non-graph-based Method

3.1.1 Latent Dirichlet Allocation (LDA)

LDA, a classic topic modeling technique, discovers hidden topics within a corpus by assigning words to topics probabilistically[2]. It uncovers hidden 'topics' within a corpus by probabilistically assigning words in documents to these topics. Each document is viewed as a mixture of topics, and each topic is characterized by a distribution of words and topic probabilities.

$$p(d|\alpha, \beta_{z_n}^v) = \int p(\theta_d|\alpha) \prod_n \sum_{z_n} p(w_{d,n}|z_n, \beta_{z_n}^v) p(z_n|\theta_d) d\theta_d$$

where β is $k \times V$ topic-word matrix. $p(w_{d,n}|z_n, \beta_{z_n}^v)$ is probability for word $w_{d,n}$ to happen when topic is z_n .

However, LDA has a limitation known as the "independence" problem. It treats words as independent and doesn't consider their order or relationships within documents. This simplification can hinder LDA's ability to capture contextual dependencies between words. To address this, models like Word2Vec and GloVe have been developed, taking word order and dependencies into account to provide more nuanced representations of textual data.

3.1.2 Latent Semantic Analysis (LSA)

LSA is a method to uncover the underlying semantic structure in textual data. It achieves this by assessing the semantic similarity between words using document-word matrices[4]. LSA's fundamental concept involves recognizing semantic connections among words based on their distribution within a document. To accomplish this, LSA relies on linear algebra techniques, particularly Singular Value Decomposition (SVD), to condense the document-word matrix into a lower-dimensional representation. This process allows semantically related words or documents to be situated in proximity within this reduced space.

$$X = U\Sigma V^T$$
$$Sim(Q, X) = R = Q^T X$$

where X is $t \times d$ matrix, a collection of d documents in a space of t dictionary terms. Q is $t \times q$ matrix, a collection of q documents in a space of t dictionary terms. U is term eigenvectors and V is document eigenvectors.

LSA, an early form of topic modeling, excels at identifying semantic similarities among words. Nonetheless, it has its limitations, particularly in its inability to fully capture contextual information and word relationships.

3.1.3 Neural Topic Model (NTM)

Traditional topic modeling has limitations, including sensitivity to initialization and challenges related to unigram topic distribution. The Neural Topic Model (NTM) bridges topic modeling and deep learning, aiming to enhance word and document representations to overcome these issues.

At its core, NTM seamlessly combines word and document representations by embedding topic modeling within a neural network framework. While preserving the probabilistic nature of topic modeling, NTMs represent words and documents as vectors, leveraging them as inputs for neural networks. This involves mapping words and documents into a shared latent space, accomplished through separate neural networks for word and document vectors, ultimately leading to the computation of the topic distribution.

The computational process of NTM includes training using back-propagation and inferring topic distribution through Bayesian methods and Gibbs sampling.

$$p(w|d) = \sum_{i=1}^K p(w|t_i)p(t_i|d)$$

where t_i is a latent topic and K is the pre-defined topic number. Let $\pi(w) = [p(w|t_1), \dots, p(w|t_K)]$ and $\theta(d) = [p(t_1|d), \dots, p(t_K|d)]$, where π is shared among the corpus and θ is document-specific. Then above equation can be represented as the vector form:

$$p(w|d) = \phi(w) \times \theta^T(d)$$

3.2 Graph-Based Methods

3.2.1 Global Random Topic Field

To capture word dependencies within a document, the graph structure incorporates topic assignment relationships among words to enhance accuracy[9].

GloVe-derived word vectors are mapped to Euclidean space, while the document’s internal graph structure, identified as the Word Graph, operates in a non-Euclidean domain. This enables the Word Graph to uncover concealed relationships that traditional Euclidean numerical data representation cannot reveal.

Calculating the "structure representing word relationships" involves employing a Global Random Field (GRF) that encodes the graph structure in the document using topic weights of words and the topic connections in the graph’s edges. The GRF formula is as follows:

$$p(G) = f_G(g) = \frac{1}{|E|} \phi(z_W) \sum_{(w', w'') \in E} \phi(z_{w'}, z_{w''})$$

The above-described Global Topic-Word Random Field (GTRF) shares similarities with the GRF. In the GTRF, the topic distribution (z) becomes a conditional distribution on θ . Learning and inferring in this model closely resemble the EM algorithm. The outcome, denoted as $p_{GTRF}(z|\theta)$, represents the probability of the graph structure considering whether neighboring words (w' and w'') are assigned to the same topic or different topics. This is expressed as:

$$p_{GTRF}(z|\theta) = \frac{1}{|E|} Multi(z_W|\theta) \times \sum_{(w', w'') \in E} (\sigma_{z_{w'} = z_{w''}} \lambda_1 + \sigma_{z_{w'} \neq z_{w''}} \lambda_2)$$

Where σ_z is a function that returns 1 if the condition x is true and 0 if x is false.

3.2.2 GraphBTM

While LDA encounters challenges related to data sparsity, particularly when modeling short texts, the Biterm Topic Model (BTM) faces limitations in its expressiveness, especially when dealing with documents containing diverse topics[13]. Additionally, BTM relies on bitwords in conjunction with the co-occurrence features of words, which restricts its suitability for modeling longer texts.

To address these limitations, the Graph-Based Biterm Topic Model (GraphBTM) was developed. GraphBTM introduces a graphical representation of bitterms and employs Graph Convolutional Networks (GCN) to extract transitive features, effectively overcoming the shortcomings associated with traditional models like LDA and BTM.

GraphBTM’s computational approach relies on Amortized Variational Inference. This method involves sampling a mini-corpus to create training instances, which are subsequently used to construct graphs and apply GCN. The inference network then estimates the topic distribution, which is vital for training the model. Notably, this approach has demonstrated the capability to achieve higher topic consistency scores compared to traditional Auto-Encoding Variational Bayes (AEVB)-based inference methods.

3.2.3 Graphical Neural Topic Model (GNTM)

LDA, in its conventional form, makes an assumption of independence. It posits that each document is generated as a blend of topics, with each topic representing a distribution over the words within the document. However, this assumption of conditional independence, also known as exchangeability, overlooks the intricate relationships and context that exist among words in a document.

The No Variational Inference (NVI) algorithm presents a departure from this independence assumption. NVI is a powerful technique for estimating the posterior distribution of latent topics in text data. It leverages a neural network structure, employing a reparameterization trick to accurately estimate the genuine posterior distribution for a wide array of distributions.

$$\alpha(\text{prior}) \rightarrow z(\text{topic}) \text{ from } \theta \rightarrow G_d(\text{structure}) \rightarrow V(\text{word set})$$

$$p(G_d^0|Z_d; M) = \prod_{(n,n') \in E_d^0} m_{z_{d,n}, z_{d,n'}} \prod_{(n,n') \notin E_d^0} (1 - m_{z_{d,n}, z_{d,n'}})$$

$$p(G_d, \theta_d, Z_d; \alpha) = p(V_d|Z_d, G_d^0)p(G_d^0|Z_d) \prod_{n=1}^{N_d} p(z_{d,n}|\theta_d)p(\theta|\alpha)$$

Unlike the Variational Autoencoder (VAE), which is primarily employed for denoising and data restoration and can be likened to an ‘encoder + decoder’ architecture, NVI serves a broader purpose and can handle a more extensive range of distributions. It’s based on the mean-field assumption and employs the Laplace approximation method, replacing challenging distributions like the Dirichlet distribution with the computationally efficient logistic normal distribution[8].

Based mean field assumption:

$$q(\theta_d, Z_d|G_d) = q(\theta_d|G_d; \mu_d, \delta_d) \prod_{n=1}^{N_d} q(z_{d,n}|G_d, w_d, n; \varphi_{d,n})$$

$$L_d = E_{q(Z_d|G_d)}[\log p(G_d^0|Z_d; M) + \log p(V_d|Z_d, G_d^0; \beta)] \\ - KL[q(\theta_d|G_d)||p(\theta_d)] - E_{q(\theta_d|G_d)} \sum_{n=1}^{N_d} KL[q(z_{d,n}|G_d, w_{d,n})||p(z_{d,n}|\theta_d)]$$

This substitution simplifies parameter estimation, making it more tractable and readily differentiable. In the context of the Global Neural Topic Model (GNTM), the logistic normal distribution facilitates the approximation of correlations between latent variables, allowing for the utilization of dependencies between topics. Additionally, the Evidence Lower Bound (ELBO) in NVI is differentiable in closed-form, enhancing its applicability.

The concept of topic proportion is represented by the equation:

$$\theta_d = \text{softmax}(N(\mu_d, \delta_d^2))$$

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(\text{logit}(x)-\mu)^2}{2\sigma^2}} \frac{1}{x(1-x)}$$

This equation encapsulates the distribution of topics within a document, reflecting the proportions of different topics in that document.

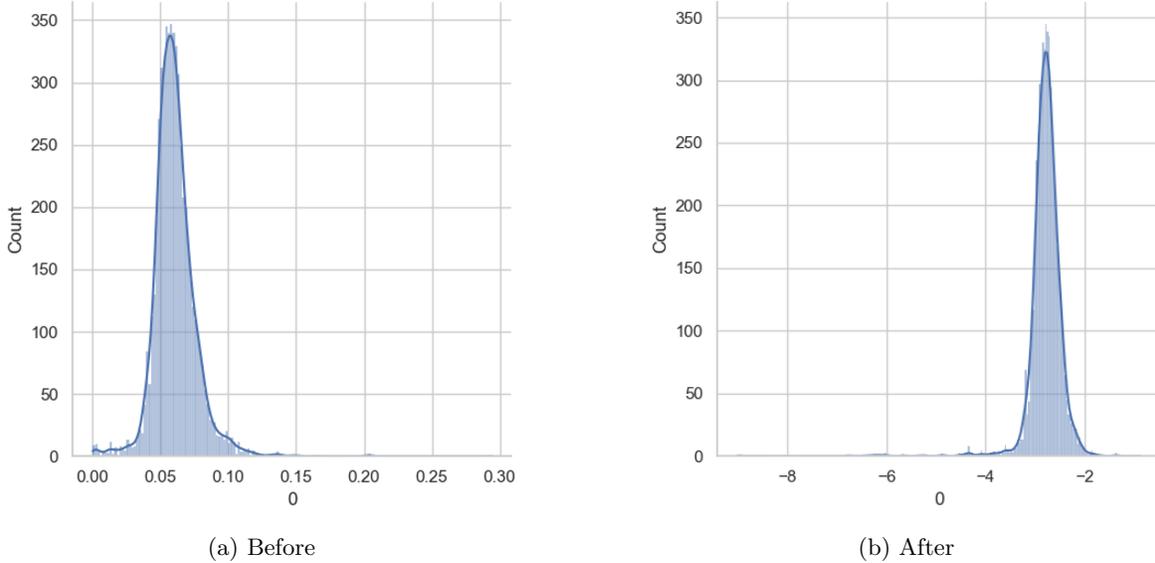


Figure 2: Transformation of logit-normal distribution after conversion

3.3 Visualization Techniques

3.3.1 Fast Unfolding of Communities in Large Networks

This algorithm aids in detecting communities within topic-words networks, facilitating interpretation and understanding of topic structures.

3.3.2 Uniform Manifold Approximation and Projection (UMAP)

UMAP is a nonlinear dimensionality reduction technique that preserves the underlying structure and patterns of high-dimensional data while efficiently visualizing it in lower dimensions. It outperforms traditional methods like t-SNE in preserving data structure.

3.3.3 Agglomerative Hierarchical Clustering

Hierarchical clustering is an algorithm that clusters data points, combining them based on their proximity until a single cluster remains. It provides a dynamic and adaptive way to maintain cluster structures, even when new data is added. Additionally, several evaluation metrics, including the Silhouette score, Calinski-Harabasz index, and Davies-Bouldin index, assist in selecting the optimal number of clusters for improved data understanding and analysis.

4 Method

4.1 Graphical Neural Topic Model(GNTM) as Factor Analysis

GNTM can be viewed from a factor analysis perspective, as it employs concepts similar to factor analysis to unveil intricate interrelationships in data and extract topics. GNTM can extract θ , which signifies the proportion of topics in each document, for summarizing and interpreting document content. In this case, θ follows a logistic normal distribution, enabling the probabilistic modeling of topic proportions.

The θ can be represented as follows[1][7]:

$$\tilde{\theta} \sim \text{LN}(\mu, \sigma^2)$$

For $0 < \tilde{x} < 1$ and $\sum_i^K x_i = 1$:

$$y = [\log(\frac{x_1}{x_D}), \dots, \log(\frac{x_{D-1}}{x_D})]^T$$

Probability Density Function (PDF) for X :

$$f_X(x; \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \frac{1}{\prod_{i=1}^K x_i(1-x_i)} e^{-\frac{1}{2} \{ \log(\frac{x}{1-x}) - \mu \} \Sigma^{-1} \{ \log(\frac{x}{1-x}) - \mu \}}$$

where the log and division in the argument are element-wise. This is due to the diagonal Jacobian matrix of the transformation with elements $\frac{1}{x_i(1-x_i)}$

GNTM shares similarities with factor analysis, which dissects complex data into factors associated with each topic to unveil the data's structure. In factor analysis, the aim is to explain observed data using latent factors. Similarly, GNTM treats topics in each document as latent variables, and these topics contribute to shaping the word distribution in the document. Consequently, GNTM decomposes documents into combinations of words and topics, offering an interpretable method for understanding document similarities and differences.

4.2 Akaike Information Criteria (AIC)

The Akaike Information Criterion (AIC) is a crucial statistical technique for model selection and comparison, evaluating the balance between a model's goodness of fit and its complexity. AIC aids in selecting the most appropriate model from a set of models.

In the context of this thesis, AIC is employed to assess the fit of a Graphical Network Topic Model (GNTM) and determine the optimal model. Since GNTMs involve parameters related to the number of topics in topic modeling, selecting the appropriate number of topics is a significant consideration. AIC assesses various GNTM models based on the choice of the number of topics and assists in identifying the most suitable number of topics.

AIC can be represented by the following formula:

$$AIC = -2 \cdot \log\text{-likelihood} + 2 \cdot \text{number of parameters}$$

Where:

- The log-likelihood is a measure of the goodness of fit of the model to explain the data.
- Number of parameters indicates the count of parameters in the model.

AIC weighs the tradeoff between a model's log-likelihood and the number of parameters, which reflects the model's complexity. Lower AIC values indicate better data fit while favoring simpler models. Therefore, the model with the lowest AIC is considered the best. AIC plays a pivotal role in enhancing the quality of topic modeling in GNTM by assisting in managing model complexity when choosing the number of topics.

For our current model, following a Logistic Normal Distribution, we utilize GNTM's likelihood:

$$L(\theta|D) = \prod_{d=1}^D \left[-\frac{1}{2} \log(|2\pi\Sigma|) - \sum_{i=k}^K (\log \theta_i - \log(1 - \theta_i)) - \frac{1}{2} \left\{ \log \left(\frac{\theta}{1-\theta} \right) - \mu \right\} \Sigma^{-1} \left\{ \log \left(\frac{\theta}{1-\theta} \right) - \mu \right\} \right]$$

When applied to a formula, it appears as:

$$AIC = -2 \cdot l(\theta) + 2 \cdot \text{number of topics}$$

Where:

$$l(\theta) = \sum_{d=1}^D \left[-\frac{1}{2} \log(|2\pi\Sigma|) - \sum_{k=1}^K \log(\theta_k(1 - \theta_k)) + -\frac{1}{2} (\log(\frac{\theta}{1-\theta}) - \mu_i)^T \Sigma^{-1} (\log(\frac{\theta}{1-\theta}) - \mu_i) \right]$$

This encapsulates the essence of GNTM and AIC in evaluating and selecting models.

5 Result

5.1 Model Setup

5.1.1 Data

The data consists of news related to the top 200 companies by market capitalization on the NASDAQ stock exchange. These news articles were collected by crawling Newsdata.io in August. Analyzing this data can provide insights into the trends and information about companies that occurred in August. Having a specific timeframe like August helps in interpreting the analysis results clearly.

To clarify the research objectives, companies with fewer than 10 articles collected were excluded from the analysis. Additionally, a maximum of 100 articles per company was considered. As a result, a total of 13,896 documents were collected, and after excluding irrelevant documents, 13,816 were used for the analysis. The data format is consistent with the "20 News Groups" dataset, and data preprocessing methods similar to those in Shen(2021)[10] were applied. This includes steps like removing stopwords, abbreviations, punctuation, tokenization, and vectorization. You can find examples of the data in the Appendix.

5.1.2 Parameters

"In our experiments, as the dataset contained a large number of words and edges, it was necessary to reduce the number of parameters for training while minimizing noise and capturing important information. To achieve this, we set the threshold for the number of words and edges to 140 and 40, respectively, which is consistent with the configuration used in the BNC dataset, a similar dataset. The experiments were conducted in an RTX3060 GPU environment using the CUDA 11.8 framework, with a batch size of 25. To determine the optimal number of topics, we calculated and compared AIC values for different numbers of topics. Based on the comparison of AIC values, we selected 20 as the final number of topics."

5.2 Evaluation

5.2.1 AIC

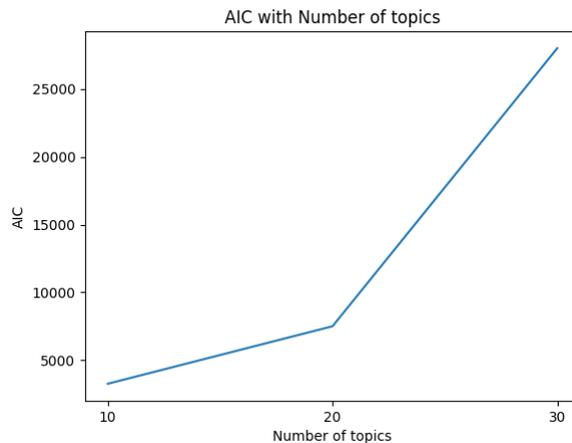


Figure 3: Changes in AIC values depending on the number of topics

AIC is used in topic modeling as a tool to select the optimal number of topics. However, AIC is a relative number and may vary for different data or models. Therefore, when using AIC to determine the optimal number of topics, it is important to consider how this metric applies to your data and model.

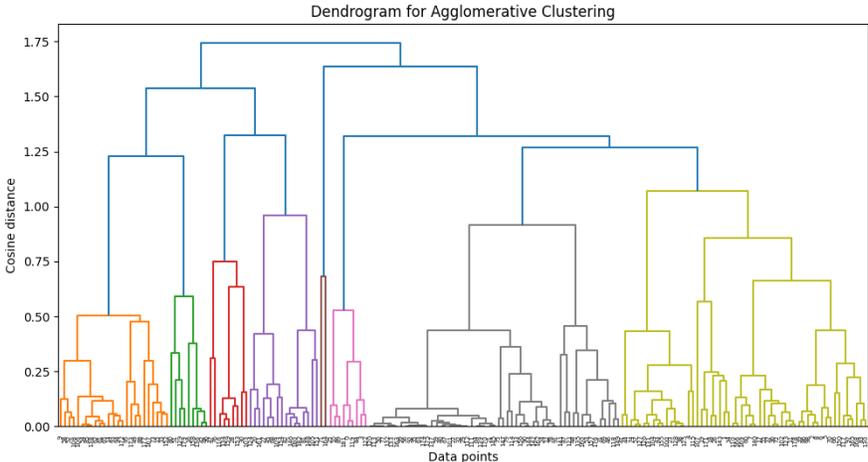
In our study, we calculated the AIC for a given dataset and model architecture and used it to select the optimal number of topics. This approach served as an important metric for finding the best

number of topics for our data. The AIC was used to evaluate the goodness of fit of our model, allowing us to compare the performance of the model for different numbers of topics.

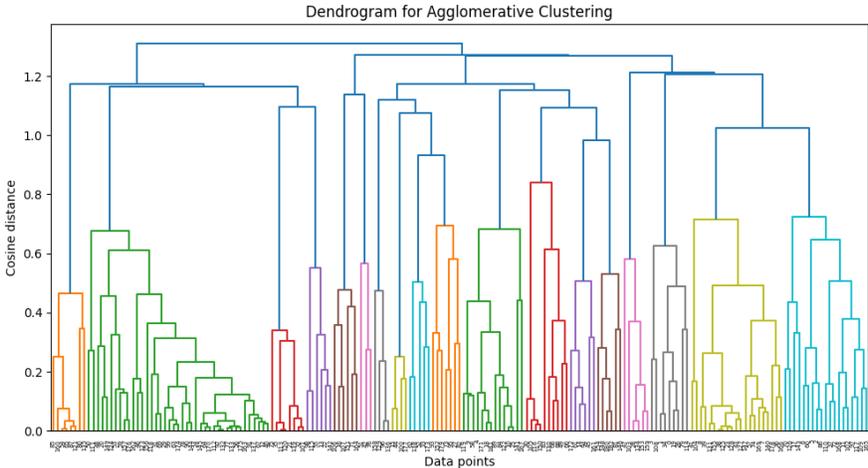
Additionally, AIC allows us to evaluate the performance of our model in comparison to AICs obtained from other models or other datasets. This allows us to determine the relative superiority of our model and highlights that we can perform optimized hyperparameter tuning for our own data and model, rather than comparing to other models. This approach is one of the key strengths of our work, contributing to a greater emphasis on the effective utilization and interpretation of topic models.

5.2.2 Topic Interpretation

5.2.3 Classification



(a) 10 Topics graph



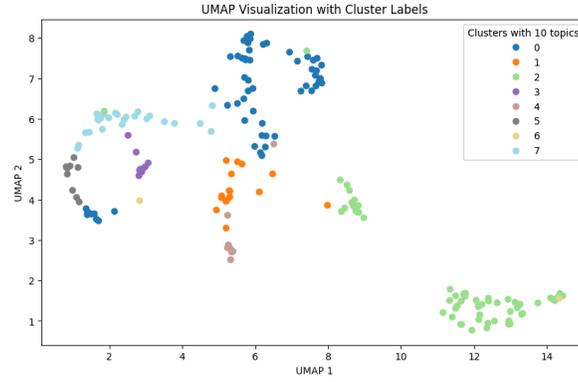
(b) 30 Topics graph

Figure 4: The result of Agglomerative Clustering

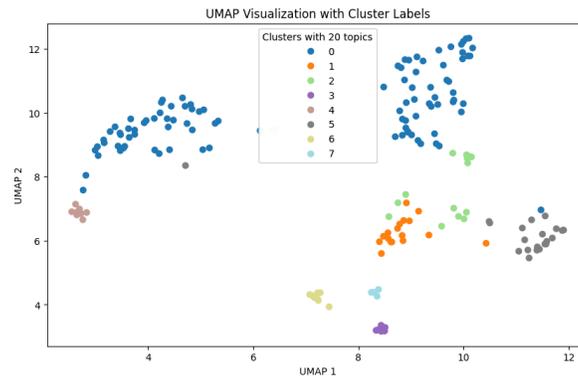
In our study, we leveraged Agglomerative Clustering and UMAP to classify and visualize news data. In our experiments, we found that news is generally better classified when the number of topics is 10. These results suggest that the model is able to group and interpret the given data more effectively.

However, when the number of topics is increased, broader topics tend to be categorized into more detailed topics. This results in news content being broken down into relatively more detailed topics, but the main themes may not be more apparent.

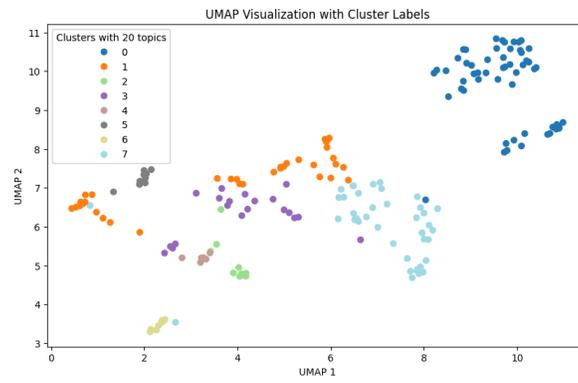
Also, as the number of topics increases, the difference in the proportion of topics that represent the nature of the news increases. This indicates a hierarchy between major and minor topics, which can



(a) UMAP graph with 10 topics



(b) UMAP graph with 20 topics



(c) UMAP graph with 30 topics

Figure 5: The result of Agglomerative Clustering

be useful when you want to fine-tune your investigation of different aspects of the news. This diversity provides important information for detailed topic analysis in context.

Therefore, when choosing the number of topics, we need to consider the balance between major and minor topics. By choosing the right number of topics, the model can best understand and interpret the given data, and we can tailor the results of the topic analysis to reflect the key features of the news content.

6 Discussion

6.1 Limitation

Even though this paper has contributed to addressing various challenges related to textual data analysis, it is essential to acknowledge some inherent limitations in the proposed methodology:

1. **Noise Edges Issue**

The modeling approach used in this paper introduces a challenge related to noise edges in the data, which can be expected when dealing with extensive corpora or numerous documents from various sources.

To effectively mitigate this noise issue, it is crucial to implement regularization techniques tailored to the specific objectives and nature of the data. Approaches such as the one proposed by Zhu et al. (2023)[12] enhanced the model’s performance by more efficiently discovering hidden topic distributions within documents.

2. **Textual Data Versatility**

While this paper focuses on extracting and utilizing the topic latent space from text data, it is worth noting that textual data analysis can have diverse applications across various fields.

In addition to hierarchical clustering, there is potential to explore alternative recommendation models, such as Matrix Factorization methods like NGCF(Neural Graph Collaborative Filtering)[11] and LightGCN(Light Graph Convolutional Network)[6], which utilize techniques like Graph Neural Networks(GNN) for enhancing recommendation performance.

Acknowledging these limitations is essential for a comprehensive understanding of the proposed methodology’s scope and areas for potential future research and improvement.

6.2 Future Work

While this study has made significant strides in addressing key challenges in the analysis of textual data and extracting valuable insights through topic modeling, there remain several avenues for future research and improvement:

1. **Enhanced Noise Handling**

The modeling used has shown promise but is not immune to noise edge issues often encountered in extensive datasets. In this study, we used a dataset comprising approximately 9,000 news articles from 194 countries, totaling around 5 million words. To mitigate these noise edge issues effectively, future work can focus on developing advanced noise reduction techniques or data preprocessing methods tailored to specific domains, further enhancing the quality of extracted topics and insights.

2. **Cross-Domain Application**

While the study showcased its effectiveness in the context of news articles, extending this approach to other domains presents an exciting opportunity. Adapting the model to different domains may require domain-specific preprocessing and feature engineering, as well as considering transfer learning approaches. Models based on Graph Neural Networks (GNN) and Matrix Factorization, such as Neural Graph Collaborative Filtering (NGCF) and LightGCN, can be employed to enhance recommendation systems and knowledge discovery in diverse fields. This cross-domain versatility can unlock new possibilities for leveraging textual data to extract meaningful insights and improve decision-making processes across various industries and research domains.”

7 Conclusion

In the context under discussion, the term "group information" pertains to the topic proportions represented by θ . From my perspective, I have undertaken an endeavor that can be characterized as Non-Linear Factor Analysis (FA) applied to textual data, analogous to traditional FA methods employed with numerical data. This undertaking proved intricate due to the inherent non-triviality in its extraction, thus warranting the classification as Non-Linear FA. (Indeed, there exists inter-topic covariance.)

Hitherto, the process has encompassed the extraction of information from textual data, a task which may appear formidable for utilization. This encompasses the structural attributes of words and topics, the proportions of topics, as well as insights into the prior distribution governing topic proportions. These constituent elements have facilitated the quantitative characterization of information within each group.

A central challenge encountered in the realm of conventional Principal Component Analysis (PCA) and FA techniques lies in the absence of definitive answers, given our inherent limitations. Consequently, the interpretation of the extracted factors poses formidable challenges and lacks assuredness. However, the GNTM methodology applied to this paper, in tandem with textual data, furnishes a network of words for each factor, thereby affording a means for expeditious interpretation. If the words assume preeminence within Topic 1, they afford a basis for interpretation. This alignment with the intentions of the GNTM. In effect, this model facilitates the observation of pivotal terms within each topic (factor) and aids in the explication of their conceptual representations.

This research has presented a comprehensive methodology for the analysis of textual data using Graphical Neural Topic Models (GNTM). The paper discussed how GNTM leverages the advantages of both topic modeling and graph-based techniques to uncover hidden patterns and structures within large text corpora. The experiments conducted demonstrated the effectiveness of GNTM in extracting meaningful topics and providing valuable insights from a dataset comprising news articles.

In conclusion, this research contributes to advancing the field of textual data analysis by providing a powerful framework for extracting interpretable topics and insights. The combination of GNTM and future enhancements is expected to continue facilitating knowledge discovery and decision-making processes across various domains.

Nevertheless, a pertinent concern arises about inordinate amount of noise pervade newspaper data or all data. Traditional methodologies employ noise mitigation techniques such as Non-Negative Matrix Factorization (NMF) and the execution of numerous epochs for the extraction of salient tokens. In the context of this research, as aforementioned, the absence of temporal constraints allowed for the execution of epochs as deemed necessary.

However, computational efficiency was bolstered through the reduction in the number of topics, while remaining the primary objectives from a clustering perspective by finding out the optimized number of topic by AIC and agglomerative clustering. This revealed that a reduction in the number of topics resulted in the observation of words associated with the original topics within sub-networks of the diminished topics.

Future research can further enhance the capabilities of GNTM by improving noise handling techniques and exploring cross-domain applications.

References

- [1] ATCHISON, J., AND SHEN, S. M. Logistic-normal distributions: Some properties and uses. *Biometrika* 67, 2 (1980), 261–272.
- [2] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
- [3] CHOI, M. J., AND KIM, K. K. Import demand in developed economies. In *Economic Analysis (Quarterly)* (2019), vol. 25, Economic Research Institute, Bank of Korea, pp. 34–65.
- [4] EVANGELOPOULOS, N. E. Latent semantic analysis. *Wiley Interdisciplinary Reviews: Cognitive Science* 4, 6 (2013), 683–692.
- [5] HAN, K. J. Analysis and implications of overseas market provision system based on domestic logistics big data. *KISDI AI Outlook 2022*, 8 (2022), 17–30.
- [6] HE, X., DENG, K., WANG, X., LI, Y., ZHANG, Y., AND WANG, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (2020), pp. 639–648.
- [7] HINDE, J. *Logistic Normal Distribution*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 754–755.
- [8] KINGMA, D. P., AND WELLING, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [9] LI, Z., WEN, S., LI, J., ZHANG, P., AND TANG, J. On modelling non-linear topical dependencies. In *Proceedings of the 31st International Conference on Machine Learning* (Beijing, China, 22–24 Jun 2014), E. P. Xing and T. Jebara, Eds., vol. 32 of *Proceedings of Machine Learning Research*, PMLR, pp. 458–466.
- [10] SHEN, D., QIN, C., WANG, C., DONG, Z., ZHU, H., AND XIONG, H. Topic modeling revisited: A document graph-based neural network perspective. *Advances in neural information processing systems* 34 (2021), 14681–14693.
- [11] WANG, X., HE, X., WANG, M., FENG, F., AND CHUA, T.-S. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (jul 2019), ACM.
- [12] ZHU, B., CAI, Y., AND REN, H. Graph neural topic model with commonsense knowledge. *Information Processing Management* 60, 2 (2023), 103215.
- [13] ZHU, Q., FENG, Z., AND LI, X. Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (2018), pp. 4663–4672.

A News Data Example

Google courts businesses with ramped up cloud AI Synopsis The internet giant unveiled new AI-powered features for data searches, online collaboration, language translation, images and more at its first annual Cloud Next conference held in-person since 2019. AP Google on Tuesday said it was weaving artificial intelligence (AI) deeper into its cloud offerings as it vies for the business of firms keen to capitalize on the technology. The internet giant unveiled new AI-powered features for data searches, online collaboration, language translation, images and more at its first annual Cloud Next conference held in-person since 2019. Elevate Your Tech Process with High-Value Skill Courses Offering College Course Website Indian School of Business ISB Product Management Visit Indian School of Business ISB Digital Marketing and Analytics Visit Indian School of Business ISB Digital Transformation Visit Indian School of Business ISB Applied Business Analytics Visit The gathering kicked off a day after OpenAI unveiled a business version of ChatGPT as tech companies seek to keep up with Microsoft , which has been ahead in powering its products with AI. "I am incredibly excited to bring so many of our customers and partners together to showcase the amazing innovations we have been working on," Google Cloud chief executive Thomas Kurian said in a blog post. Most companies seeking to adopt AI must turn to the cloud giants – including Microsoft, AWS and Google – for the heavy duty computing needs. Those companies in turn partner up with AI developers – as is the case of a major tie-up between Microsoft and ChatGPT creator OpenAI – or have developed their own models, as is the case for Google.